Doctoral theses at NTNU, 2023:111

Parisa Rezaee Borj

Online Grooming Detection on Social Media Platforms

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of Science and Technology

Parisa Rezaee Borj

Online Grooming Detection on Social Media Platforms

Thesis for the Degree of Philosophiae Doctor

Gjøvik, April 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

ISBN 978-82-326-6484-9 (printed ver.)
ISBN 978-82-326-5587-8 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:111

Printed by NTNU Grafisk senter

Declaration of Authorship

I, Parisa Rezaee Borj, hereby declare that this thesis and the work presented in it are entirely my own. Where I have consulted the work of others, this is always clearly stated. Neither this nor a similar work has been presented to an examination committee elsewhere.

signature:

.....

(Parisa Rezaee Borj)

Gjøvik, Date: April 2023

Abstract

Online grooming detection has become a critical research topic in the era of extensive data analysis. It is essential to protect vulnerable users, particularly adolescents, against sexual predation on online platforms and media. However, many factors challenge online grooming detection, which leads to a high-risk problem for youth. The primary goal of this research work is to provide techniques that increase children's security on online chat platforms. To this extent, many experiments have been conducted to create models fulfilling our research goal. As such, this thesis contains a comprehensive survey of child exploitation in chat logs that provides the readers with a deep knowledge of the problem, possible research gaps, and proposed solutions. In this research, we split the online grooming detection problem into several subproblems, including author profiling, predatory conversation detection, predatory identification, and data limitations issues.

The leading theory behind the author profiling in this problem comes from the fact that online predators provide fake identities to tarp their young victims. At the same time, children's characteristics differ from the ones who imitate a minor, which leads us to detect the gender of users in this research. In this thesis, we propose a gender detection model that can recognize the gender of authors based on their keystroke dynamics features. This research also provides a fake identity detection technique with a high performance that detects users who are dishonest about their identity.

Providing an automatic predatory conversation detection system facilitates law enforcement authorities to act on time before any tragedy occurs. Therefore, we have examined and proposed several predatory conversation detection and predatory identification techniques focusing on finding the best feature vectors and embeddings that lead to the best performance in online grooming detection.

This thesis also aims to gain deep knowledge about predatory behaviour with se-

mantic analysis. We might lose some semantic information by applying conventional embeddings such as Word2vec or GloVe feature vectors since they provide a single word embedding for a term in different contexts. At the same time, humans show their motivations in phrases or sentences rather than single terms. So, we provide an online grooming detection model based on extracting embeddings from sentences rather than single words. We apply contextual model based such as Bert-based and RoBerta-based systems for each sentence.

Several constraints, such as privacy and security issues, availability, and the imbalanced nature of the datasets, challenge online grooming datasets. The number of predatory chat logs is considerably lower than the other online conversations, leading to a highly imbalanced data problem. It is challenging to build a machine learning model based on imbalanced datasets, which motivates us to provide a model to handle this issue. This research proposes a model that uses a hybrid sampling and class re-distribution to gain augmented data for coping with highly imbalanced datasets. We also improve the diversity of classifiers and feature vectors by perturbing the data along with the augmentation in an iterative manner.

Finally, we conclude our research by discussing potential research gaps and open problems and proposing possible solutions for them to give deep insights to the readers of future work based on the work of this thesis.

Acknowledgement

I would like to thank my supervisor Professor Patrick Bours for providing me with an incredible opportunity to pursue Ph.D. I am extremely grateful for your trust in my abilities and the freedom that allowed me to explore my scientific interests. Also, I would like to thank my co-supervisor Kiran Raja, for his constant support over the duration of this thesis, given the fact that my research work would not be gone so far without his technical support, and patience. I thank the administration of the department of information security and communication technology, NTNU, Gjøvik, for providing me with a good research work environment. I would like to express my gratitude to the dissertation committee members for spending their time reviewing my thesis and evaluation. I would like to thank Kyle Porter, who showed me that any difficult situation can be handled when someone makes life fun, enjoyable, and full of meaning. Ultimately, I would like to thank my mother, friends (Ahmad Hassanpour, Bahare Elhami, and Mazaher Kianpour), colleagues, and family members for their encouragement and support.

Contents

List of Tables

List of Figures

List of Abbreviations

AB	AdaBoost
Acc	Accuracy
AP	Author Profiling
BNC	British National Corpus3
BoW	Bag of Words
CBoW	Continuous Bag of word
CSAM	Child Sexually Abused Material
EM	Expectation-Maximization
FN	False Negative
FP	False Positive
GNB	Gaussian Naive Bayes
KD	Keystroke Dynamics
KNN	K-Nearest Neighbor
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
ME	Maximum Entropy
PJ	Perverted Justice

PP Press Press Latency

PR Press Release Latency

Pre Precision

Rec Recall

RF Random Forest

RP Release Press latency

RR Release Release Latency

SVM Support Vector Machine

TC Term Count

TF Term Frequency

TF-IDF Term Frequency Inverse Document Frequency

TN True Negative

TP True Positive

Part I

Overview

Chapter 1

Introduction

The internet has changed children's lives through new digital technologies. They can access diverse sources of information on different online platforms, such as online gaming and social media, along with the chance of online communication and new friendships. At the same time, children can easily be exposed to inappropriate, or potentially harmful content, such as being manipulated for sexual abuse [3, 4]. New friendships and romantic relationships using the internet as a medium is inevitable for most adolescents. This phenomenon causes much harm, including unwanted sexual relationships, minor exploitation and abuse, and online grooming. Online media has been used for exchanging child pornography, finding potential victims, engaging in a dangerous relationships, and normalizing certain destructive behaviours to lower the child's inhibition. In recent years, many cases have been reported on misuse of online media for sexual abuse of children [5, 6]. Many research works in digital forensics have tried to detect sexual abuse on minors occurring online [4, 7, 8, 9, 10, 11]. Significant amount of works have focused on tracking and detecting distribution methods for images and videos [7, 8, 9]. Few other works have focused on policy and legal dimensions [4, 7, 10, 11].

Some sexual predators might only go through available online child pornography, while others prefer to get in touch with a minor in real life gradually starting with messages [1]. Even if a victim has no physical contact, the damage of such an act can impact a child's life tremendously. In this thesis, we mainly focus on cases where child predators use different applications such as chat rooms and social network applications (e.g., Twitter, Facebook, and Instagram) with the help of text messages to engage in a relationship with minors. To succeed in convincing a minor to provide sexual favors, a predator typically grooms a victim. Thus, it is not only important to detect the predatory activity, but also to detect the grooming process at an early stage to mitigate any subsequent harm. In some cases, such

as public chats, these types of risks can be mitigated by employing a moderator, while in private communications, using a moderator is not possible and makes the problem challenging [1]. It is not only a complex problem to monitor a private peer-to-peer conversation with respect to privacy issues but also needs tedious efforts considering the enormous scale of this type of communication in a large peer-to-peer social media interaction [12]. Therefore, we mainly focus on peer-to-peer chat conversations to detect the sexual abuse of minors on online platforms in an automated manner. Our primary goal is, therefore, to analyze a scenario where a predator (i.e., a malicious actor intending to get sexual favors from minors) targets a victim using text messaging as the primary mode.

1.1 Motivation and Problem Statement

Internet-based child predation has become a focus of online grooming research and legislation. One can consider online grooming as a procedure of gaining, persuading, and engaging a minor age user in sexual activity, where the internet is used as a medium for conversation. Some reports show that many underage users have been the target of online grooming behaviour [5, 6]. A report by the National Center for Missing and Exploited Children[1] showed more than 16 million child abuse cases in 2019. A retrospective report in the United States noted that around 65 percent of users who chatted with an unknown adult as a minor had experienced some sexual solicitation [5]. 23% of the minors had an online conversation with a stranger adult that followed a grooming pattern. Among the ones who had long online relationships with predators, less than half (38%) met the predators in person. At the same time, many of those encountered in reality reported physical and sexual abuse [5]. In 2019, 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 the New York Times claimed that this number doubled in only one year [6].

In 2021, a report by the Internet Organised Crime Threat Assessment (IOCTA)[2] showed that online grooming activities on online platforms have increased drastically. One should also consider that the new encrypted messaging platforms can ease the distribution of child sexual abuse material (CSAM) materials between predators, where self-generated material is one of the primary threats against adolescents. Also in 2021, a report by End Child Prostitution and Trafficking (ECPAT)[3] showed that the COVID-19 pandemic has led to a higher demand for digital lives, increasing the chance of children being targeted by online offenders. The increas-

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

[2] https://www.europol.europa.eu/publications-events/main-reports/iocta-report

[3] https://ecpat.org/

ing number of these reports leads to a concern that requires attention.

One of the significant detriments of online media platforms is the possibility of hiding real identities by users. Knowing online users' demographic attributes has become a vast field of research that early research works call author profiling [13]. There are various reasons of interest for author profiling, such as marketing, forensics, or security [14]. Also, many molesters provide fake identities to trap their young victims and build a relationship with them for further abuse [14]. The hypothesis behind author profiling is that adults who impersonate minors to trap victims may have a set of different characteristics compared to actual minors. For example, they may use different linguistic patterns when they share information in online conversations compared to children. So, addressing the grooming detection problem from the author's profiling perspective allows for identifying possible impersonators among social media users for early detection of child grooming, preventing further peril.

Early research papers considered online interaction between a predator and the victim as cyber grooming [15, 16, 17]. The grooming procedure by the predator is not always under a fake identity. Instead, many predators use their actual age or identity but inform their victims that their relationship is unsuitable for lowering future punishment in case of detection [1]. We consider online grooming detection as a multi-faced aspect that can be automated in different manners. Besides looking into the demographic attributes of the users, we can also focus on the conversation to analyze a grooming pattern. The main idea behind predatory conversation detection is that predatory conversations have a different pattern from non-predatory ones. Also, finding potential predators who send suspicious messages can facilitate grooming detection in the early stages. Machine learning models can detect these differences, find suspicious chat logs, and identify predators.

The primary motivation of this thesis is to detect cyber grooming by machine learning models. Considering the threat posed by online predators, developing reliable grooming detection techniques are essential. Therefore, the main goal of this doctoral work is to develop reliable and automated models that increase minors' security on online platforms. It should also be mentioned that detecting online activities in front of a camera or live stream that produce self-generated CSAM materials requires video and image processing which is out of the scope of this thesis. Therefore, we limited the thesis's scope to text-based approaches.

1.2 Ethical-Legal Issues

From the ethical-legal perspective, applying the term "predator" can be challenging and objectional. The term "predator" can offend one's dignity and the pre-

sumption of innocence, while it is not yet proven if the person is an actual child molester. This thesis uses the term "predator" since it is the official word for online grooming detection in computer science research areas. However, in the case of other research areas or contexts, it may be advisable to use the term "presumed predator" or "potential predator" to avoid ethical-legal complications.

1.3 Research Objectives

Given the above motivation, the objectives of this research work are to:

- Understand the state-of-the-art by performing a comprehensive survey of child grooming and identify the challenges in detecting child exploitation for sexual favors in the literature.

- Investigate if the authorship profiling techniques can help in grooming detection. One of our objectives is also to understand the role of age and gender on the behavioural patterns of online users.

- Develop novel techniques for predatory detection, especially for predatory conversation detection and predatory identification in two-users chat logs. We aim to benchmark predatory conversation detection techniques and analyze different feature extraction approaches to gain better detection accuracy in predatory conversation detection.

- Gain a deeper understanding of predatory patterns through semantic analysis. We aim to investigate different approaches to understand the semantics of sentences rather than term entities, as phrases and sentences can reflect human behaviour.

- Develop approaches in a data-imbalance setting as the data for grooming detection is not publicly available due to privacy issues and regulations. The fraction of grooming messages is marginally small compared to all online conversations. One of our primary purposes in this research is to create an approach that can provide good performance despite an imbalanced dataset in grooming detection.

1.4 Research Questions

The following research questions are formulated to be addressed in this thesis for online grooming detection purposes:

1.4.1 Research Question 1 (RQ1): Author Profiling

Anonymity in online platforms can challenge user authentications for security purposes. Online predators might fake their age and gender to attract their minor victims. One of the main steps for avoiding any risks of threatening minor victims is finding risky cyber grooming situations by detecting potential predators who lie about their demographic attributes. The adolescents might believe that the predator has the same age as their own, seeing his deceptive behaviour and leads to harmful acts by the adult predator [18]. So, the first questions are:

1. Can we detect the users who fake their identities on online chat platforms?
   (a) How can we authenticate users' profiles correctly based on Keystroke Dynamics (KD)?
   (b) Can KD information discover if a user tries to mimic the behaviour of the other gender or another age group?
2. Can we determine the demographic characteristics of users based on their behaviour in a chat room?
   (a) Can the typing pattern reveal the author's age group and gender?
   (b) What kind of features (stylometry-based/typing rhythm-based (KD)/ensemble) can detect the age and gender of a user in chat logs?

1.4.2 Research Question 2 (RQ2): Predatory Conversation Detection

Automatic predatory conversation detection in an enormous amount of chat data helps law enforcement authorities with early detection. However, it is a complex problem to detect grooming before physical abuse since predators have different tactics concerning their personalities and fear of being detected. It has been shown that ordinary people like parents or school teachers cannot detect grooming behaviour since the predatory conversation might not include any topic that indicates inappropriate relationship or sexual abuse [19]. The reasons above have motivated this research to focus on predatory conversation detection, where online grooming messages can be distinguished from ordinary ones.

1. Can we detect a predatory conversation from an ordinary chatlog?
   (a) What is the best technique to detect predatory conversation reliably?
   (b) Does preprocessing the chatlogs result in performance gain for predatory conversation detection? How much preprocessing must be performed in mining chat room conversations to detect online grooming?

(c) Which features are more discriminative for detecting the grooming chat lines and predatory conversations?

1.4.3 Research Question 3 (RQ3): Predatory Identification

It is essential to identify the predators in a chat room for forensics. Following the same motivation, in this research, we have tried to identify the suspicious users in chat logs, and the main question is:

1. Can we identify predators from the normal users in chat conversations?
   (a) What text features give the best performance in detecting predators?

1.4.4 Research Question 4 (RQ4): Data Constraints

Several data constraints, such as availability, privacy issues, non-standard structure, and the unreliability of online data, challenge grooming detection. In this thesis, we have mainly focused on two limitations of datasets in grooming detection problems: imbalanced data problems and transferability of detection approaches in cross-domain settings.

Imbalanced Data Issue: The number of grooming messages publicly available is approximately only 0.25% of all messages on online media [20]. The small proportion of predatory v/s non-predatory conversations leads to imbalanced data problems where machine learning techniques can give unreliable results. Applying a balanced dataset for training a machine learning model is critical, particularly in cases where the result will be used in a real-life scenario, such as a court-of-law that can permanently impact a person's life. We, therefore, investigate data imbalance by answering the following questions.

1. Can we create a predatory detection model that deals with heavy imbalanced data to arrive at reliable decisions?
   (a) Is it possible to re-distribute the grooming data, so it becomes balanced?
   (b) What kind of re-sampling can produce a promising performance for grooming conversation detection?

Transferability of Detection Approaches in Cross-Domain Settings: Different domains have different context-specific expressions and terms that have different meanings for the new domain's context. Predatory conversations do not follow the same pattern as natural language. Depending on the predator's personality

and the chat's condition, the language theme in grooming messages vary. Child molesters do not explicitly show their intentions, making detecting a grooming conversation complicated. A deep understanding of predatory behaviour utilizing semantic analysis can reveal the motives behind a conversation. We, therefore, pose the main question to investigate the transferability of detection approaches as below.

1. Can a predatory detection technique be developed by semantic analysis transferability in cross-domain settings and result in reliable decisions?
   (a) What type of semantic analysis can be used in cyber grooming detection?
   (b) Which features can distinguish predatory behaviour in different semantic contexts?

1.5 Research Methodology

The research methodology is given below to achieve the research goals based on the earlier-mentioned research questions. The thesis uses extensive empirical research where the methodology involves creating necessary datasets to answer the questions formulated. We start by conducting a comprehensive survey of the state-of-the-art and then develop approaches for different sub-problems.

- **State-of-the-art Survey: A survey for an overall view of the problem**
  A comprehensive survey of child exploitation in chat logs is performed. To give a better overview of the problem, we divided the online grooming detection problem into sub-topics with a detailed review. The survey presents a detailed analysis of earlier works for each sub-problem and compares them to show their strength and weaknesses. Also, the data and feature vectors for grooming detection are explained, along with their constraints and limitations. It also discusses future research directions with open research questions.
- **New Dataset: Keystroke Dynamics datasets**
  There is a lack of diverse data for authorship profiling by keystroke dynamics to detect users' attributes such as age and gender. For gender detection by keystroke dynamics information, we collected data where all participants chatted remotely via Skype and could freely choose the topic of the conversations as it happens in a real-life scenario. As mentioned above, one of the main concerns in chat room security is ensuring that the person behind an online profile is the one he/she claims to be. As such, we conducted another

data collection that we could use for fake identity detection. In data collection for liar detection, participants chatted in two scenarios: with a real identity and the fake identity.

  - **Gender Detection: based on Keystroke Dynamics features**
    It is vital to warn a person that the gender of a chat partner profile is fake. For instance, when a male subject impersonates a female to get close for further abuse, it is critical to warn as soon as possible before any harm occurs. Following this motivation, we proposed a model that can detect chatters' gender by applying keystroke dynamics and stylometry feature sets.
  - **Fake Profile Detection: based on Keystroke Dynamics features**
    As mentioned earlier, we conducted a data collection where participants could use a fake identity to chat with others. We analyzed this data, mainly the keystroke dynamics information, to reveal dishonest ones about their online identity.

- **Predatory Conversation Detection Approaches**
  From the forensics point of view, it is essential to have an automated system that separates the suspicious chat logs from the non-predatory ones. It will decrease the time consumed by police to analyze the related conversations for conducting an action on time before any harm happens. So, we propose automated models that differentiate the grooming conversations from normal ones. We also discuss and show what suitable preprocessing and feature extractions should be done to improve performance in predatory conversation detection.

- **Predatory Identification Approach**
  Sometimes molesters are engaged in several chat logs seeking inappropriate relationships with minor adolescents. It might be more beneficial to identify suspicious users than to detect suspicious conversations since it might reduce the time spent. Therefore, we also present a model that identifies predators based on their behaviour in chat rooms.

- **Semantic Transferability of Detection Approaches in Cross-Domain Settings**
  Distributed representations such as word2vec and GloVe vectors encode the semantics of linguistic units, such as words or word-like entities. They provide one single-word embedding for each term in different contexts. Although, deeper language processing models require a more profound knowledge of a human language to reveal the semantics behind the words. Phrases

and sentences represent human knowledge or common sense [21], showing that internal sentences display semantic intuitions [22, 23]. Prior knowledge of a word or entity is defined in the context of a sentence or a phrase. Many applications such as machine translation, image captioning, and dialogue systems need to understand the semantics of sentences rather than word-like entities to improve their performance in understanding human languages [23]. In this research, we have used a simple contrastive sentence embedding framework (SimCSE) which uses a contrastive objective with pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) or an improved BERT model called RoBERTa.

- **Approach for Handling the Imbalanced data**
  We provide a new approach that applies a new strategy for handling the imbalanced dataset in grooming detection. The proposed approach creates a balanced class distribution by increasing the minor class (predatory class) with a set of augmented and perturbed data.

1.6 Major Contributions

This section gives an overview of our main contributions to this research work.

1.6.1 A Comprehensive Survey

An extensive survey of existing literature on child exploitation is presented to provide deep knowledge regarding each research question and possible solutions, along with the potential constraints. The survey is complemented with psychological theories and their use in machine learning models for grooming detection problems. The survey categorizes all available datasets and approaches, and their constraints in predatory detection, along with research gaps and possible future solutions [1].

1.6.2 Author Profiling

We contribute a new dataset and approach for gender detection [24] and fake identity detection [25] using keystroke information while composing text. The proposed gender detection approach provides a promising performance of 72% by analyzing free-text data captured in 15 minutes. Further, we contribute an approach for fake identity detection based on a single message with an accuracy of more than 70% and correct classification of a whole chat with well over 90% accuracy. This contribution part fulfills Research Question 1 (RQ1).

1.6.3 Predatory Conversation Detection and Predatory Identification

We present four different approaches for predatory conversation detection, and predatory identification [26, 27, 28, 29] as listed below.

1. **Pre-Processing techniques**: We proposed better preprocessing to gain better detection, and identification accuracy than existing state-of-art algorithms [28]. The proposed preprocessing approach fulfills RQ2 and RQ3.

2. **Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles**: We contribute a new approach for handling the imbalanced dataset in predatory conversation detection [27]. Our proposed approach creates a balanced class distribution by increasing the number of predatory samples (minor class) with a set of augmented and perturbed data (until a 50% balance is obtained) to gain better accuracy over state-of-the-art approaches. The proposed approach for handling the imbalanced dataset fulfills RQ2 and RQ4.

3. **Online Grooming By Simple Contrastive Chat Embeddings**: We contribute a new approach based on a contrastive learning framework for feature extraction from sentences and conversations with misspellings. The approach takes into consideration not only the phrases in the message, but also the semantics, and we demonstrate it to provide good detection accuracy on public datasets [29]. The proposed approach in this section fulfills RQ2 and RQ4.

1.7 List of Research Publications

This section provides a list of publications as a result of research conducted during the doctoral study.

- Borj, P. R., Raja, K., & Bours, P. (2022). Online grooming detection: A comprehensive survey of child exploitation in chat logs. Knowledge-Based Systems, 110039.

- Borj, P. R., & Bours, P. (2019, May). Detecting liars in chats using keystroke dynamics. In Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications (pp. 1-6).

- Li, G., Borj, P. R., Bergeron, L., & Bours, P. (2019, May). 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) (pp. 1049-1054). IEEE.

- Borj, P. R., & Bours, P. (2019, October). Predatory conversation detection. In 2019 International Conference on Cyber Security for Emerging Technologies (CSET) (pp. 1-6). IEEE.
- Borj, P. R., Raja, K., & Bours, P. (2020, October). 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 (pp. 1-6). IEEE.
- Borj, P. R., Raja, K., & Bours, P. (2021, September). Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles. In 2021 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.
- Borj, P. R., Raja, K., & Bours, P (2023). Detecting Online Grooming By Simple Contrastive Chat Embeddings, 9th ACM International Workshop on Security and Privacy Analytics (IWSPA 2023) [Accepted].

1.8 Thesis Outline

Figure 1.1 represents the overview of the tasks carried out during this doctoral research work. In part I, Chapter 1 describes the introduction and real-life incidents that have motivated this work. It also represents the problem statement and the research objectives. We have detailed the research questions in Section 1.4 to address in this thesis, followed by the research methodology in Section 1.5. Our main contributions are summarized in Section 1.6.

Chapter 2 presents the background of grooming detection problem and existing literature. Further, we describe the evaluation metrics that measure the performance of grooming detection models (see Section 2.4). Furthermore, a summary of the research articles published in this doctoral thesis is presented in Chapter 3. Chapter 4 gives a summary of the proposed solutions for each research question discussed in Section 1.4, and finally, in part I, Chapter 5 discusses the constraints that limit online grooming detection in real-life scenarios along with the future works and open research gaps based on this thesis.

To solve the online grooming problem, we divided the problem into three main subproblems: author profiling, predatory detection, and data constraints. For each subproblem, we have proposed techniques listed in the papers in part II. Part II

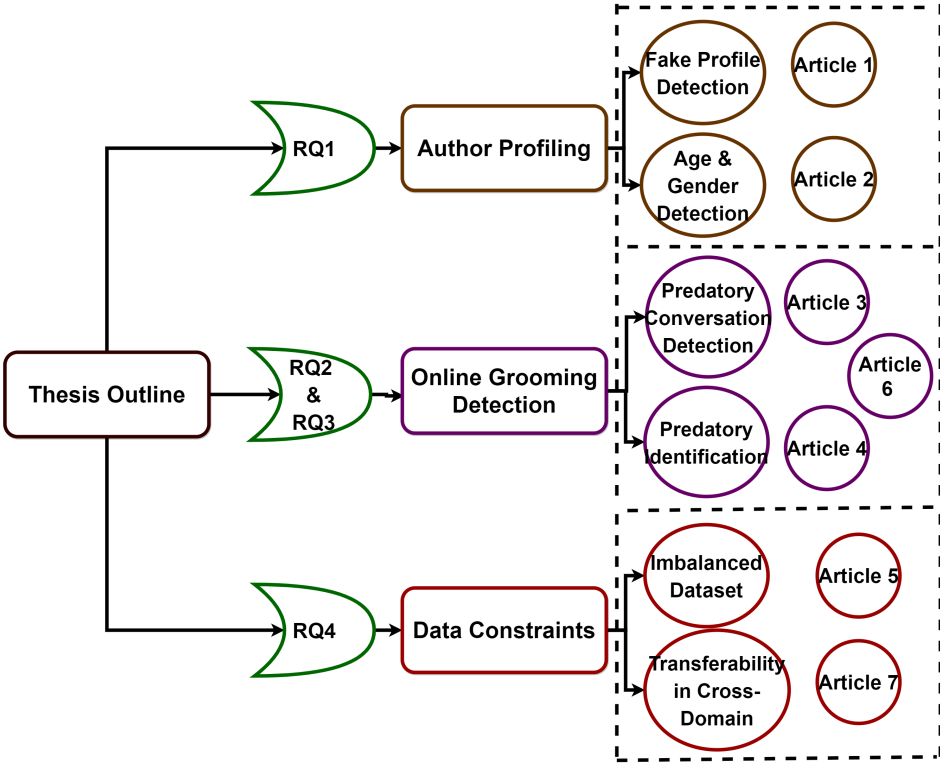

Figure 1.1: Overall structure of the thesis

starts with surveying all existing literature for child exploitation detection in chat logs (see Chapter 6). In part II, Chapter 7 represents our research paper that proposes a model for handling fake identity detection. Chapter 8 presents our paper on the gender detection model on online platforms. Chapter 9 and Chapter 10 present the proposed models for predatory detection and predatory identification along with the proposed method for suitable pre-processing techniques in grooming detection problems. Chapter 11 represents our work on handling the imbalanced nature of datasets in online grooming detection, and finally, Chapter 12 presents our approach for grooming detection by contrastive chat embeddings.

Chapter 2

Related Work

The prevention of online sexual violence by applying machine learning techniques has become a critical topic [1, 30, 31]. This chapter summarizes our recent survey article (Chapter 6) that details the existing research on child exploitation on online chat platforms [1]. We have created an overview of various methods researchers applied to cope with online grooming problems. It covers grooming definitions, feature extraction techniques, machine learning models, and evaluation metrics.

2.1 Background of Online Grooming

- **Online Grooming:** Grooming refers to the situation where the predator prepares the environment and the child for further abuse [15]. Some research works [32, 33] used the term pedophile for the predator. At the same time, it should be considered that pedophilia is a very determined clinical diagnosis that cannot be applied to all offenders [15]. The intention of the different predators varies from one case to another [34]. Some might like to see the victim in reality (contact child sex offender (CCSO)), while others are just fantasy-driven and do not try to meet the minor victim in the real world (fantasy child sex offender (FCSO)) [34, 35]. An early research work [35] investigated the contextual difference of contract-driven predators from others who are only curious or fantasy-driven. In both cases, predators attempt to build a trusted relationship. In a CCSO case, offenders might use tactics to convince the minor to meet offline, while FCSO cases do not [35]. However, the damage led by both types of grooming to victims is severe and should be avoided [1].

- **Online Grooming Stages:** During the process of cyber-grooming, several interactions are carried out, first to gain the victim's confidence and after

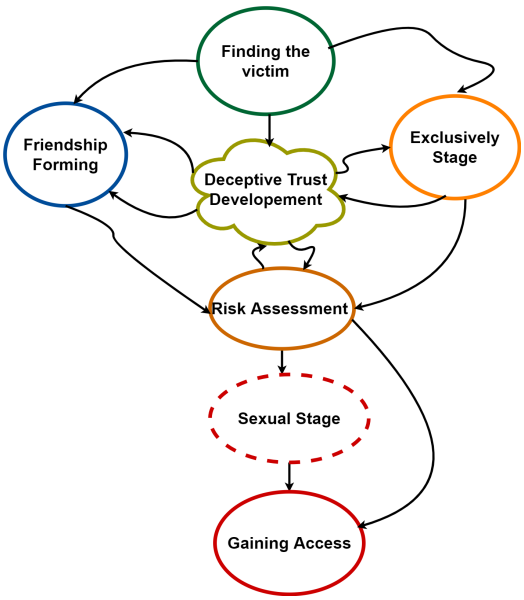

Figure 2.1: Various Stages in a Grooming Procedure [1]

that, to abuse them with malicious intents [15]. The predators manipulate the child and the people around the victim to abuse the child without being caught [19]. The main goal of their strategy is to avoid detection by being charming and helpful towards the victim. Their manipulative behaviour contains particular strategies such as finding a suitable victim, gaining access to the victim, developing a confidential and trusted relationship, and finally desensitizing the victim for further abuse [16, 17]. Researchers have categorized the grooming procedure into different stages while the predator slowly prepares the minor to gain his/her trust [19, 36, 37, 38]. It starts with victim selection and friendship forming, followed by the trust development stage, risk assessment stage, exclusivity stage, sexual stage, and conclusion stage. It should be mentioned that victim selection is the first step or stage of the grooming procedure that depends on many factors, such as attractiveness and vulnerabilities [36, 39]. Finding the easy victim, the predator attempts to develop a trusted relationship in the following grooming stages while assessing the various risks [37]. In some cases, the offender talks about his previous relationships with sexual topics. In many cases, the offender avoids sexual topics not to scare the minor victim and tries to get in touch with him/her in the conclusion part. The predators might not sequentially follow the grooming stages. Their primary goal is to reduce the chance of losing the victims' trust [40]. So, they might change their plans based on

the situation and return to the early stages to avoid any disclosure. Figure 2.1 displays different grooming stages in a predatory conversation and how the transition from one state can go back to the previous step. For instance, some predators do not mention sexual topics in a conversation not to scare the minor user and then loos his/her trust [1].

2.2 Data

The data is prominent in designing a proper module for online grooming detection. It is essential that the data follows the exact nature of the real-world data where messages are short or full of misspellings and slang words. The nature of the public posts on social networks such as Twitter and Facebook are different from a private conversations. However, considering the privacy issues, the lack of data in this field led the researchers to use data from Twitter or blogs to develop a grooming detection model [1]. The data used in this field is summarized in Table 2.1:

Data	Sources	Ref
Predatory Data	www.perverted-justice.com	[16, 17, 18, 30, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48]
	PAN2012	[28, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63]
	MovieStarPlanet	[64]
Non-Predatory Data	www.literotika.com	[41, 42, 44]
	http://www.irclog.org	[30]
	http://krijnhoetmer.nl/irc-logs	[30]
	Omegle	[30]
	Twitter	[14, 65, 66, 67, 68, 69, 70]
	Blogs, Book Reviews	[18, 47, 71]
	British National Corpus(BNC)	[72, 73]

Table 2.1: A summary of datasets used for grooming detection in previous works [1]

2.2.1 Data Constraints

The amount of predatory conversation is considerably low compared to the regular discussions on all online platforms. So, the data used for online grooming detection should follow the same nature [30]. Designing a machine learning model based on biased data with good performance detecting both classes is challenging [27]. Similarly, many online annotations or labels on online platforms are fake. It is vital that researchers evaluate the data collection process and make sure the labels are valid [1]. The constraints of data for cyber grooming detection are summarized below:

- Non-Standard Structure
- Security and privacy issues
- Imbalanced Nature

- Ethical considerations
- Non-reliable labels and self annotators

We also should consider that some factors, such as the context or time of communication, represent excellent knowledge about the authors and their intentions. So, applying multi-modal data facilitates risk detection by increasing the performance of the grooming detection model [1, 31].

2.3 Features

The primary hypothesis behind grooming detection by machine learning models is that grooming conversations have different characteristics (features) from non-predatory messages. The characteristics can differentiate between classes of persons (e.g., adults versus children), or between the themes of the conversation or text (e.g., predatory versus non-predatory messages). Previous works have applied different feature vectors to capture different characteristics of the predators for grooming detection [24, 28, 51, 52, 53, 61, 62, 63].

2.3.1 Textual Features

The majority of research work applied textual or lexical features for predatory detection. Most papers used a one-hot representation of text feature vectors such as different types of the bag of words, including binary, term frequency, term count, and TF-IDF [16, 26, 28, 47, 48, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 61, 62, 63, 64]. The one-hot representations of the words have multiple constraints, such as the sparsity of the feature vectors and the lack of distinguishing the meaning of the words used in different contexts. The BoW feature vectors do not contain information about the word's meaning based on different contexts and locations. They overlook the analogies of the terms that come from different meanings in different texts. A distributed representation of the data considers the meanings of the words in different conversations. Following the same motivation, some early research works have applied GloVe and Word2vec as the distributed representations of the chat terms [28, 27, 53, 58].

Considering the fact that predators show feigned emotion and affection to make the impression that they are in love with the minor victim, it is critical to capture the psychological characteristics by the words that can reveal the affective features of a conversation. They try to develop trusted relationship by showing fake emotions and controlling their victims for further harm. Some research works have categorized the psycholinguistic profiles for the conversations revealing the emotional and psychological aspects of the data as Linguistic Inquiry and Word Count (LIWC) [74, 75]. For instance, Parpar et al.[60] have categorized 80 types of LIWC fea-

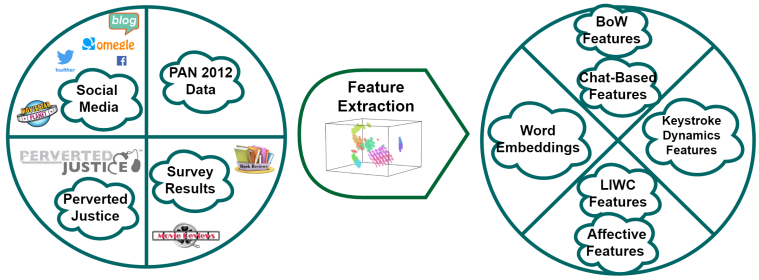

Figure 2.2: Data and Features in Sexual Predatory Detection [1]

tures for predatory detection in chatrooms. Also, chat-based features, 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, can give good hints for detecting a suspicious conversation. [31, 60, 76]. For instance, the change of the relationship over time for assessing the threat was captured by Elzinga et al. [77]. Figure 2.2 displays a summary of the discussed data and features.

2.3.2 Keystroke Dynamics Features

Keystroke Dynamics is a way of user authentication or identification based on the rhythm of their typing on the keyboard [78]. For instance, checking the correctness of a password can include both information like what the password is and how the password was typed on the keyboard. One of the first examples of applying the KD information is when the telegraph operators could identify the other by their Morse code rhythm of typing ("The Fist of the Sender" used during World War II) [79]. Various software captures the KD information, such as BioPassword[1] and BeLT [80], where BioPassword is used for password hardening.

Keystroke dynamics is a low-cost method to detect users' demographic features such as age and gender. Many early research works applied it for age, and gender detection [81, 82, 83, 84, 24, 25]. A summary of various keystroke dynamics features is detailed below:

1. **Keycode feature:** Keycode is the ASCII value of each pressed key on the keyboard.
2. **Duration feature:** The duration is the time that a key remained pressed. In some research papers is considered as dwell-time or hold-time.
3. **Latency feature:** The latency shows the time between releasing a key and pressing the next one.

[1] http://www.biopassword.com/

4. **Press-Press(PP)-latency feature:** The PP-latency is the time between pressing one key and the following key.

5. **Release-Release(RR)-latency feature:** The RR-latency time shows as the time interval between successive key releases.

2.3.3 Feature Constraints

Several limitations can challenge the discriminative and stability of the feature vectors in grooming detection limitations. For instance, BoW features provide extremely sparse feature vectors that downgrade the machine learning model's performance. Also, predator conversations have many overlaps with non-predatory conversations. It causes a lack of boundary in various datasets classes and uncertainty in different feature vectors. For example, [62] tried to reduce the feature space by applying the fuzzy-rough method to select the most discriminative features for grooming detection. One should also consider the limitation of applying pre-trained networks such as Word2vec and GloVe based on Google News data which is unsuitable for sexual predatory detection problems due to the different nature of terms in conversations.

2.4 Evaluation Metrics

Considering the predatory samples for classification as positive samples and the rest negative, the grooming detection is a two-class classification problem. True Positive (TP) samples are the predatory samples that have been classified correctly as positive cases. True Negative (TN) samples are the non-predatory samples that have been correctly classified as non-predatory cases. A predatory sample classified as negative is considered a False Negative (FN) classification. A False Positive (FP) is a negative sample classified as a predatory case.

This research evaluated models' performance by applying standard metrics 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.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

- **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.

$$P = \frac{TP}{TP + FP} \tag{2.2}$$

Precision indicates the probability that a sample that is classified as predatory is in fact predatory.

- **Recall** is the fraction between detected relevant samples (TP) and all the actual relevant samples, i.e.

$$R = \frac{TP}{TP + FN} \tag{2.3}$$

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

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \tag{2.4}$$

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

Applying accuracy as the only scale for measuring the performance in grooming detection techniques can cause some challenges where the data is extremely imbalanced. For instance, in the case where less than one percent of the dataset is predatory samples, not correctly classifying the positive samples (predatory samples) can still lead to high accuracy in providing incorrect performance reporting.

Early research works used the F_β score for compensating the accuracy flaw [30]. When precision and recall are given equal weights, the β is set to 1 and F_1-score is used. When there is a focus on minimizing false positives than minimizing false negatives, a $\beta < 1$ is employed. When the priority is to minimize false negatives, a $\beta > 1$ is employed. In our thesis, we focus on minimizing the false positives and therefore use $\beta < 1$. We specifically use $\beta = 0.5$ to report performance of all the approaches in this work. The main motivation behind reporting $F_{0.5}$-score is to provide a metrics to reduce overload on law enforcement authorities by minimizing false positives (i.e., predatory conversations identified as non-predatory conversations). Further, the metric is also inline with many early research works that have given less weight to precision than recall and applied $F_{0.5}$-score as metrics for measuring online grooming performance [1, 28, 30, 50, 52, 53, 55, 61].

2.5 Online Grooming Detection Methods

Grooming detection has been investigated in different manners. We have categorized the grooming methods into four main categories. Researchers coped with the grooming problem from different perspectives, including predatory stage detection, predatory conversation detection, predatory identification, and author profiling for online grooming detection. It should also be mentioned that many investigations on the dark web consider the distribution of nude images or abused child videos [4, 7, 10, 11], which is out of this thesis's scope.

We detail the four main categories of grooming detection below:

1. **Predatory stage detection:** Various grooming characteristics have been extracted to detect suspicious conversation stages leading to grooming [17, 38, 40, 41, 42]. The characteristics represent the purpose of each grooming phase and can vary from one case to another. For instance, in the friendship-forming stage, the offender and the minor victim might exchange personal information such as email address, age, and gender. Similarly, conversations about a favorite hobby or giving compliments can indicate the trust-development phase. The overall overview of the grooming characteristics that different papers [17, 38, 40, 41, 42] have used is presented in Table 2.2. Different researchers considered different stages of cyber grooming. For example, some analyzed it as a three steps procedure [38, 45, 46, 49] while others considered it a four to six stages process [17, 40, 41, 42, 85].

2. **Predatory conversation detection:** The similarities between predatory conversations and ordinary ones have complicated grooming detection. The difficulty of detecting grooming behaviour by ordinary people such as parents, teachers, or other people around a minor victim has compelled an automated predatory conversation detection system. The enormous number of chatlog samples on online platforms also raises the complexity of this problem. So, it has motivated the researchers to find methods that prune the searching space by dividing all conversations into two main classes, including suspicious conversations and ordinary ones [25, 47, 50, 51, 52, 53, 54, 55, 58, 63, 86, 87].

3. **Predatory identification:** Some papers also identified the predators along with detecting suspicious conversations [16, 18, 57, 60, 64, 88]. The result of predatory identification can be used for forensics purposes in case of need in an upcoming court case.

4. **Author profiling for online grooming detection:** Many 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 Favourite 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

Table 2.2: Grooming stages and their characteristics [1]

Author	Number of Stages	Data	Features	ML Technique
Kontostathis [38]	3	www.perverted-justice.com	Groomng Characteristics	K-mean Decision Tree
Kontostathis et al. [45]	3	www.perverted-justice.com	Grooming Characteristics	Phrase matching Rule-based techniques
Escalante et al. [49]	3	PAN2012	BoW	Ring-based Classifier Decision Tree
Michalopoulos et al. [46]	3	www.perverted-justice.com	TF-IDF	Naive Bayes SVM Maximum Entropy EM and EMSIMPLE Decision Tree
Cano et al. [40]	4	www.perverted-justice.com	BoW, POS, sentiment, complexity, readability, length, psycho-linguistic (LIWC dimensions)	Decision Tree SVM
Gunawan et al. [41]	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. [17]	6	www.perverted-justice.com	LIWC to create psycho-linguistic profile based on grooming characteristics	Logistic Regression Decision Tree

Table 2.3: Online Grooming Stage Detection by different papers [1]

have essential roles in the writing style, such as age and gender [65, 71, 72, 73, 81, 89, 90, 91, 92]. Knowing demographic information of users, such

as age and gender, can ease child exploitation detection when the data is enormous. The amount of online data that needs to be processed is massive. Therefore, finding a method that reduces the data is critical to minimizing the search space for predatory detection. Author profiling facilitates finding offenders from adolescents by detecting gender and, more importantly, age. Researchers have used both textual keystroke dynamics features to reveal the demographic attributes of online users [81, 82, 83, 84, 89, 90].

Table 2.3 gives an overview of different research works for predatory stage detection and table 2.4 categorizes the grooming detection models into three subjets including predatory conversation detection, predatory identification, and predatory stage detection.

Objective		
Grooming Stage Detection	**Predatory Conversation Detection**	**Sexual Predatory Identification**
Cano et al. [40]	Bogdanova et al. [47]	Ashcroft et al. [18]
Egan et al. [93]	Borj et al. [28]	Borj et al. [28]
Escalante et al. [49]	Bours & Kulsrud [50]	Cardei & Rebedea [51]
Gunawan et al. [41]	Cardei & Rebedea [51]	Cheong et al. [64]
Gupta et al. [17]	Ebrahimi et al. [53]	Fauzi & Bours [55]
Kontostathis [38]	Ebrahimi et al. [52]	Misra et al. [56]
Kontostathis et al. [45]	Escalante et al. [54]	Morris [57]
Michalopoulos & Mavridis [46]	Fauzi& Bours [55]	Pendar [16]
Pranoto et al. [42]	Miah et al. [48]	Villatoro et al. [61]
	Misra et al. [56]	
	Munoz et al. [58]	
	Zuo et al. [62]	
	Zuo et al. [63]	

Table 2.4: Categorization of works according to objectives in predator detection [1].

2.6 Conclusion

This chapter is a summary of our recent survey paper (Chapter 6) [1]. It summarizes various aspects of grooming detection in literature focusing on chat conversations, starting by explaining the definition of grooming. Further, multiple phases of online grooming are defined along with their different characteristics used by machine learning models for grooming stage detection. This chapter details different sets of features applied for online predatory detection along with their constraints. We also show available datasets and their limitations. Finally, we categorize online grooming detection methods into four subsets: predatory stage detection, predatory identification, predatory conversation detection, and author profiling.

Chapter 3

Summary of Published Articles

This chapter summarizes the research articles published throughout this PhD program. The following sections present a brief overview of each article with an introduction, motivation, and research findings. The topics shown in Figure 1.1 and research questions in Section 1.4 correspond to the papers are discussed in this chapter.

3.1 Article 1: Online Grooming Detection: A Comprehensive Survey of Child Exploitation in Chat Logs

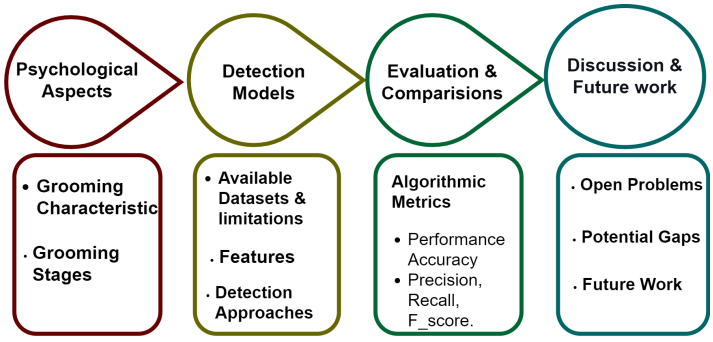

Figure 3.1: Overall contributions of survey paper [1]

Due to the lack of algorithmic surveys for grooming detection on online chat platforms, we have provided a research paper that analyzes and summarizes the state of the art for this problem with a focus on chat messages. Our survey paper focuses on analyzing all research related to peer-to-peer chat communication for sexual abuse of minors on online platforms. It considers the scenarios where a malicious actor intends to target minor victims to get sexual favors from them [1].

The psychological theories of the child grooming procedure have been investigated, and we have shown how machine learning models apply these theories to discover grooming behaviour on online platforms. This research provides an in-depth explanation of feature sets, their constraints, and potential solutions. Also, the available datasets are detailed with their limitations to supplement the readers with the state-of-the-art. We also discuss the essential role of authorship profiling and how it can facilitate the early detection of online grooming by applying keystroke dynamics and textual features. We conclude our survey paper by detailing the limitations that challenge predatory detection, along with open problems and potential solutions and future work [1]. Figure 3.1 represents the primary contributions of our survey paper.

3.2 Article 2: Detecting Liars in Chats using Keystroke Dynamics

Anonymity on online platforms provides the chance to apply fake identities for users, which can challenge the security of the online environment. Different reasons are behind the fake identities, such as collecting information for further spear phishing attacks with various intentions or impersonating a charismatic person to develop a relationship with minors. Early research has shown that deception is common on online platforms such as chat rooms, discussion boards, and online dating websites [94, 95, 96]. In general, users are also quite suspicious about the honesty of others on the internet [95, 96]. However, a minor user might not be aware of this nature of social media and be trapped by predators who fake their age and gender.

The faint and unreliable lie indicators challenge deception detection in chat rooms. However, psychological research shows that lying is more cognitively demanding as providing fake information requires more memorization to avoid mistakes. Consequently, the liar needs more time to think and process the conversations, which increases the memory load and reaction time [97, 98, 99, 100]. Early research studied response latency or Reaction Time (RT) as a behavioural feature [101]. Capturing the time while chatting can indicate some hints if the user is lying. Lying can affect emotional arousal and cognitive load and lead to behavioural change in typing [97, 102]. The main goal of this paper is to investigate the possibility of detecting liars by looking at the timing features of typing. We use keystroke dynamics information to investigate the impact of cognitive lie theory on typing behaviour.

The hypothesis behind our data collection is based on the fact that deception increases the cognitive load. So, we designed two sets of conditions; in the first one, the participant was asked to chat based on his/her true identity. While in the

second condition, we provided a fake identity for participants of different ages and gender, and they were asked to chat online based on their new fake identity. The KD information of the users was collected using BeLT[1]. To avoid any possible device-specific variation, we used an identical set of tools, such as the same laptop and the same version of Skype. We extracted all KD features for data analysis, including duration, latency, PP, and RR. Also, the message length, pause time, Backspace features, Shift features, tri-graphs features, and enter features were extracted.

We defined two scenarios for detecting fake identities. In the first scenario, every message was classified as a lie or a fact. While in the second scenario, the whole conversation was classified as a fake or a fact. In the first scenario, we used 70% of the messages for training, and the designed model was tested for single message classification by the 30% of the data. The SVM gained the best accuracy for message classification with a value of over 70%.

In the second scenario, the SVM model was used for classifying each message in a chat. The classification of each of the separate messages in the chat was used to decide if the whole chat was based on a fake identity in different manners as below:

- **Majority Classification Voting (MCV):** In this case the majority of the classifications of the messages determined the classification of the chat.
- **Sum Score Classification (SSC):** if the sum of all classification scores was negative, the whole conversation was detected as a lie and vice versa.
- **Limited Sum Score Classification (LSSC):** it was assumed that the scores close to zero value are not strongly indicating any class. We only considered the scores below or above an absolute threshold value δ.

We could detect the conversations based on fake identities with an accuracy of over 90%.

3.3 Article 3: Exploring Keystroke Dynamics and Stylometry Features for Gender Prediction on Chat Data

Following the same motivation for detecting the users with fake demographic attributes, we analyzed stylometry and keystroke dynamics features to detect the gender of the users. For this paper, we collected the data where participants chatted

[1] A Behavioural Logging Tool that has been developed at the Norwegian Information Security Laboratory (NISlab) in 2013 as part of a bachelor student project [80].

together online for around 15 minutes. We used BeLT while participants chatted remotely via Skype to capture the keystroke dynamics features. The data consists of forty-five participants contained, 35 males and ten females. We ensured the classifier was trained on a balanced dataset to avoid an imbalanced dataset. It means that during the classifier creation phase, the number of KD features selected from male and female subjects are equal, which led to highly imbalanced test sets [24].

Two sets of feature vectors are extracted based on keystroke dynamics and stylometry. The stylometry features are:

- **Average thinking time** is the time between two messages. It measures the time between releasing the last key for one message and when the user presses the first key for the following message.
- **The ratio value of key deletion** is based on all deleted keys used, divided by the total number of keys typed by a particular user.
- **The average number of letters in a word**
- **The average number of words in a message**

In addition to the above-mentioned features, we extracted the latency, duration, PP, PR, and RR.

The gender prediction model has two parts: the first part classifies the gender of the author of each message, and the second part merges the results of all message classifications to decide the gender of the author of the conversation by using the majority voting technique. The best results for gender prediction achieved by the Random Forest classifier are 72% accuracy when it only uses keystroke dynamics features and 84% accuracy when the stylometry features are used. It should be mentioned that in the training section, ten male and ten female conversations were used to handle the imbalanced nature of the data. The results showed that the pausing time between two messages and typo correction feature vectors are solid indicators for distinguishing a male from a female [24].

3.4 Article 4: Predatory Conversation Detection

Aside from author profiling, early research work also focused on chat conversations to detect predatory conversations. The predator might not hide his or her actual age or gender. In some cases, the offender even informs the minor that their relationship is not suitable to avoid further issues in potential courts and gain more trust. It is critical to look at the conversation rather than trying to detect the age or

gender of the users. Many predatory conversations are not easy to detect since the conversation seems normal, while the conversation's trajectory might give some hints. It is challenging for people who do not know the grooming tactics to recognize a predatory conversation [19]. Early research showed that ordinary people could not distinguish a writing pattern of a child predator from a normal one, and providing hints would bias their decisions [19].

The main goal of this paper was to propose a model that recognized predatory conversation from ordinary chat messages. In order to gain our goal, we analyzed the predatory conversations to see the various aspects of a suspicious conversation. Molesters are careful in using words not to scare the minors, so they do not use hardcore words while building a confidential relationship. Also, predators primarily define the topic of the conversation in order to gain more information and assess the situation. So, they ask many questions that might not be seen in an ordinary chatlog. They identify their victims' needs and give many compliments about different topics such as appearance, ability, and personality.

For our experiment, we used the data collected for PAN 2012 [30]. To avoid any bias and cover all different types of online conversations, the data includes three different subsets, such as usual (i.e., non-sexual) chat conversations, sexual conversations between consenting adults (from OMEGLE), predatory conversations from PJ. We extracted three different n-grams using the TF-IDF weighting scheme to train the classifiers. It was shown that SVM gained the best performance with an accuracy of 98%. In text analysis, there are many features, and SVM is a suitable classification method to handle data sets with large sparse feature spaces, and more precisely, linear SVM is a good approach for data analysis with high dimensional feature sets.

3.5 Article 5: On Preprocessing the Data for Improving Sexual Predator Detection

Predatory detection is a critical challenge. One can consider the problem of predatory detection as a classical two-class problem where the predators are considered in one class and the rest in another. It leads to a severe imbalanced problem that requires a robust technique. Early research work has focused on three main goals, including:

- Identification of predatory chat lines;
- Classification of predatory chat conversations;
- Identification of the predator and the victim in a predatory conversation.

In this paper, our primary goal is to investigate the textual features alone for predatory conversation identification and predator identification. Figure 3.2 represents the pipeline for predatory conversation identification and predator identification in this research work.

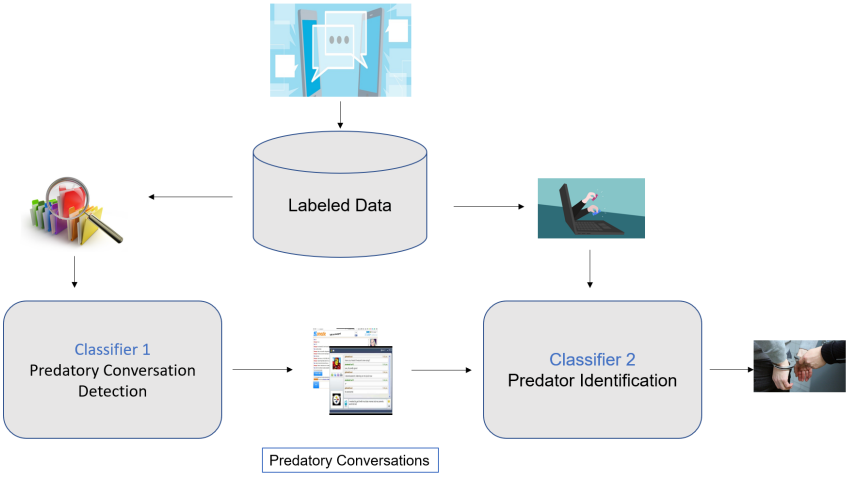

Figure 3.2: The Architecture of the Sexual Predator Detection System

We used PAN 2012 [30] data set for all experiments and evaluations. The data contains training and testing sets in a disjoint manner. Each set has many conversations, where a unique conversation ID identifies each conversation. Each message in each conversation also has an author ID, time, and the text of the message. The dataset also includes two files of author IDs for training and testing sets that contains the IDs of the predators [30].

One of the purposes of this paper is to show the impact of preprocessing on the dataset and investigate if it can improve the performance of predatory detection. The IDs of the predators were used to label the conversation into two different classes, i.e., predatory and non-predatory. We only analyzed two author messages since a real grooming conversation contains only two users. A conversation containing less than seven messages does not give enough information to classify. So, they were removed. We did not perform any stemming or lemmatization not to lose any information. However, we removed many non-meaningful words and symbols.

There is a lack of investigation on the robustness of feature space for predatory conversation detection. In this paper, we also investigate various feature vectors for predatory detection. We converted the dataset to vector representations of nu-

meric values by applying different approaches such as Bag of Words and word embeddings. Bag of Words contains all the possible words in the dictionary resulting in a high dimensional vector. There are different techniques for calculating the non-zero entries in the feature vector, such as binary, term count, term frequency, and TF-IDF. In this paper, we have extracted all mentioned features for our investigation. Since Bag of Words features do not consider the relationship between the words and lose some valuable information, we also extracted word embeddings feature sets such as GloVe by applying a pre-trained GloVe embedding model to obtain a feature vector of dimension 300.

To detect predatory conversations, we merged all the messages of a single conversation into a single text block and extracted the features from each of the merged texts. Various machine learning techniques were applied, and we illustrated the improvement in detecting predatory conversation with an accuracy of 0.994 and an F_1-score of 0.964 by training SVM with TF-IDF feature vectors. Also, the GloVe vector as feature space with SVM has given an accuracy of 0.989 and an F_1-score of 0.930 to detect predatory conversations.

Predatory identification was done using the best performing feature set and classification algorithm from predatory conversation detection. All known predatory conversations were extracted and split into two texts. One was composed by merging all the messages from the sexual predator, while the other text was composed by merging all the victim's messages, and the feature vectors were made based on each group of messages. The best-achieved result for predatory identification had an accuracy of 0.905 and an F_1-score of 0.914.

3.6 Article 6: Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles

The need for relevant data for grooming detection results in a highly imbalanced dataset challenge since the number of predatory conversation data is much less than ordinary conversations [20, 51, 62]. The grooming conversations are around 0.25% of all queries, and the state- of-the-art early research works employ a standard dataset with less than 10% predatory conversations [27]. Training the machine learning models with imbalanced data leads to a sub-optimal classifier that overlooks the small class by giving more weight to the other class and results in overfitting or underfitting [30]. The mentioned problem is more critical when the data has a skewed distribution with features such as class overlapping, small sample size, and small disjuncts. Considering the nature of grooming conversations, they overlap and are disjunct with non-grooming conversations where predators talk about ordinary topics while building a trusted relationship [103].

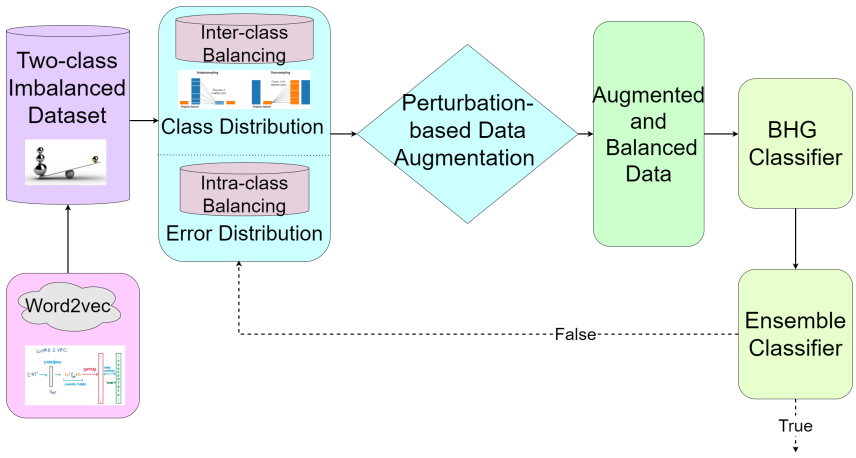

Figure 3.3: Proposed approach for predatory chat detection

We present a new approach for dealing with class imbalance using a hybrid sampling and class re-distribution to obtain an augmented dataset. Most early research works focused on conventional methods for grooming detection without considering the imbalanced nature of the grooming data [28, 30, 47, 52, 61, 76, 93]. Cardei et al. [51] applied various techniques for handling imbalanced problem such as cost-sensitive technique and sampling techniques including BalanceCascade [104], and CBO - a clustering-based method using k-means [105]. Also, an adaptive fuzzy method with artificial neural networks (ANNs) for addressing the imbalanced data in sexual grooming detection was proposed by Zuo et al. [63]. The overall pipeline of the proposed approach is illustrated in the Figure 3.3.

We present an approach that first creates a balanced class distribution by increasing the minor class (predatory conversation class) with a set of augmented and perturbed data. The balanced class distribution is increased until a 50% balance is obtained by simply augmenting and perturbing the data. With the refined class distribution, we create an ensemble of HistogramBoostedGradient classifiers, which directly benefit from the augmented and perturbed data in selecting different features for creating ensembles. With the set of experimental validation, we evaluate the proposed approach on the PAN 2012 [30] dataset, where the proposed approach outperforms the existing approaches. The proposed approach results in a precision of 99%, a recall of 99% and a $F_{0.5}$ score of 94% with a gain of 3% over the recent work which reported a recall of 96% [27].

3.7 Article 7: Detecting Online Grooming By Simple Contrastive Chat Embeddings

The main goal of a safe online environment for children is to have a robust system that knows the predatory pattern deeply. This profound knowledge of grooming behaviour will help find the predators before harm occurs. However, it is challenging to know their behaviour since it depends on many factors, from predators' personalities to time, location, and the context of chat conversations. Semantic analysis is a domain-dependent task. It can facilitate the grooming detection problem, where it can cover the context-dependency of the words in different times and situations. For instance, the same message in different contexts can give different impressions. One can consider this sentence "it can take all your time" in different situations. If this sentence describes an online website while ordering an item, it gives a negative feeling. However, in a situation where it explains an idea about a book or show series, it indicates positive impressions. Humans show their common sense and knowledge with sentences and phrases rather than a single term or word [21, 22, 23]. At the same time, traditional word embeddings such as word2vec and GloVe provide one feature vector for a word in a different context or sentence. Words have different meanings in different sentences, so it is critical to apply a semantic analysis model that considers the context-dependency in sentences for extracting the semantic characteristics. Bidirectional Encoder Representations from Transformers (BERT) or RoBERTa has profoundly bidirectional contextualization and allows the model to gain information from various representations with different positions [106].

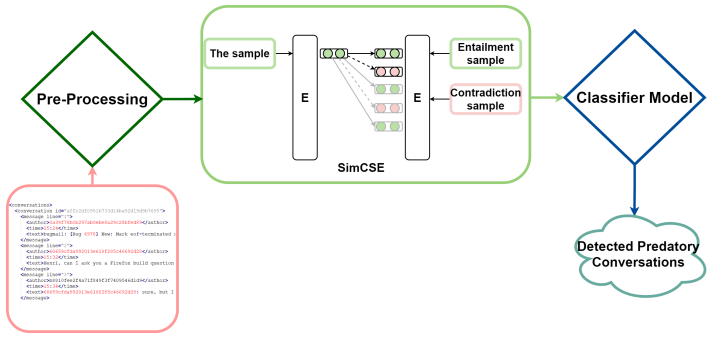

Figure 3.4: Applied SimCSE Pre-trained network for Semantic Analysis in Online Grooming Detection

To handle the difficulty of context dependency in online grooming detection, this research paper provides a contrastive learning framework for feature extraction

from the sentences. It can also assign a feature vector to the conversation with misspellings using subword information. We propose an approach with different components, including the pre-processing phase, the pre-trained network for extracting the embeddings phase, and the classifier phase (see Figure 3.4). We clean the datasets and pass them to the network to extract the sentence embeddings for the classification of the main data. The output of the pre-trained network are embeddings for sentences and are passed to the classifier models in the next component. We evaluated the proposed model on PAN 2012 [30], where the proposed approach outperforms the existing approaches. Our proposed model is a configuration of RoBERTa encoders and supervised SimCSE for training the SVM model, leading to a high rate of detecting relevant samples (predatory samples). Our proposed approach gains an F_0-score of 0.96, an F_1-score of 0.96, and an accuracy of 0.99 for predatory conversation detection that benchmarks the state-of-the-art. We also show an improvement in performance by applying different fusion models in order to gain a more reliable decision for predatory conversation detection. The sum fusion of all configurations gives an accuracy of 0.99, an F_1-score of 0.97, and an $F_{0.5}$-score of 0.98.

Chapter 4

Conclusions

This thesis aimed at developing reliable models that detect online grooming on online platforms. This thesis has three main parts, i.e., author profiling based on keystroke dynamics characteristics, online grooming detection based on text analysis, and data constraints in predatory detection. We have defined various research questions for each part and have extended much research works in terms of proposed models to address the main research questions detailed in section 1.4. The research questions investigated resulted in various publications listed in parts II. First, we have surveyed all research works on child exploitation problems, mainly chat logs, to gain a better overview of the online grooming detection problem [1] and unaddressed problems. Our survey paper presented a systematic review of the existing state-of-the-art in online grooming problems. It looks into the psychological aspect behind child grooming and how various research works have applied these aspects to define grooming characteristics for automated detection by machine learning models along with available datasets and applied feature vectors [1]. The following sections summarize the proposed solutions for each research question.

4.1 Research Question 1 (RQ1): Author Profiling

1. **Can we detect the users who fake their identities on online chat platforms?**

   (a) How can we authenticate users' profiles correctly based on Keystroke Dynamics (KD)?

   (b) Can KD information discover if a user tries to mimic the behaviour of the other gender or another age group?

   Conclusion:

Due to the unavailability of public datasets to answer this research question, we collected a new dataset that contains keystroke dynamics information when one is composing a text as self and as an assumed person in a messaging scenario. The dataset was designed with the primary goal of detecting fake identities and gender identification. To this extent, the data was gathered in two different scenarios where participants chatted with a real identity and the fake identity. The hypothesis behind deleting fake identities is that users require more time to reply as a different identity causing subtle differences in typing behaviour patterns. The results show that typing patterns can differentiate users who impersonate a fake identity from the actual ones. The performance improves significantly when multiple messages are combined to decide if the user is lying about his/her identity. We can conclude that various typing features, such as making corrections, typing speed, and pause times, are solid, distinctive characteristics for fake identity detection on social media [25].

2. **Can we determine the demographic characteristics of users based on their behaviour in a chat room?**

   (a) Can the typing pattern reveal the author's age group and gender?

   (b) What kind of features (stylometry-based/typing rhythm-based (KD)/ensemble) can detect the age and gender of a user in chat logs?

Conclusion:
To cope with the problem of detecting the demographic characteristics of users with no publicly available dataset, we conducted another data collection that covers the keystroke dynamics characteristics of authors for gender detection. In this thesis, a gender detection technique is proposed that detects the gender of authors based on their keystroke dynamics characteristics with a reasonable accuracy that is increased by using stylometry features. The results show that the pausing time between the messages and the misspelling correction features are the most discriminative characteristics for gender detection. It should be mentioned that gender prediction can be applied to warn the users that the chat partner's gender is incorrect and that he/she is trying to impersonate the other gender [24].

4.2 Research Question 2 (RQ2): Predatory Conversation Detection

1. **Can we detect a predatory conversation from an ordinary chatlog?**

   (a) What is the best technique to detect predatory conversation reliably?

(b) Does preprocessing the chatlogs result in performance gain for predatory conversation detection? How much preprocessing must be performed in mining chat room conversations to detect online grooming?

(c) Which features are more discriminative for detecting the grooming chat lines and predatory conversations?

Conclusion:
We proposed several automated models that detect the grooming conversations from the normal ones. The first primary goal was to devise an appropriate preprocessing method for short text analysis without losing valuable information [28]. As such, we proposed a preprocessing model that significantly increased predatory conversation detection performance that can be easily adapted to existing state-of-the-art approaches. Along with the preprocessing model, several feature extraction techniques were analyzed. The thesis not only used the traditional Bag of Words feature sets but also showed how applying distributed representation of words such as GloVe can significantly improve predatory conversation detection performance. To systematically demonstrate our proposed model, we analyzed them on the PAN 2012 dataset [30]. It was shown that the GloVe feature vector as feature space with SVM results in a promising performance for predatory conversation detection. Our proposed models for predatory conversation detection have performed best compared to the benchmark in the literature [26, 28].

4.3 Research Question 3 (RQ3): Predatory Identification

1. **Can we identify the predators from the normal users in chat conversations?**

   (a) What text features give the best performance in detecting predators?

Conclusion:
Our proposed model contains two phases; in the first phase, the predatory conversations are detected and in the second phase, the detected predatory conversations are used to distinguish a victim from a predator. For predatory identification, we applied the results from the best-performing feature vectors and classification algorithms for predatory conversation detection. To train the predatory identification model in the second phase, each conversation was split into two text blocks containing the victims' and predator's messages. Therefore, we classified the users of detected predatory conversations into two classes; predator versus victim. The results show that one can distinguish a predator from a victim in a conversation with an accuracy of 0.905 and an F_1-score of 0.914 [28].

4.4 Research Question 4 (RQ4): Data Constraints

Several data constraints challenge cyber grooming detection problems, such as availability, privacy issues, non-standard structure, and the unreliability of online data. This research mainly focused on two types of data limitations: imbalanced data problems and transferability of detection approaches in cross-domain settings. This thesis answers the questions below:

1. **Can we create a predatory detection model that deals with heavy imbalanced data to arrive at reliable decisions?**

   (a) Is it possible to re-distribute the grooming data, so it becomes balanced?

   (b) What kind of re-sampling can produce a promising performance for grooming conversation detection?

   Conclusion:
   The number of predatory conversations on online platforms is considerably lower than regular conversations. So, one should consider the issues of such heavy imbalanced datasets for developing a reliable predatory detection model. Therefore, this work has proposed a new approach for handling the imbalanced nature of predatory data by hybrid sampling and class redistribution to obtain an augmented dataset. Also, to improve the diversity of classifiers and features in the ensembles, our model proposed to perturb the data along with augmentation in an iterative manner. We have gained an improvement of 3% over the best state-of-the-art approach, an F_1-score of 0.99, and an $F_{0.5}$ of 0.94 [27].

2. **Can a predatory detection technique be developed by semantic analysis transferability in cross-domain settings and result in reliable decisions?**

   (a) What type of semantic analysis can be used in cyber grooming detection?

   (b) Which features can distinguish predatory behaviour in different semantic contexts?

   Conclusion:
   Since predators conceal the primary motivation behind their problematic behaviour, it is critical to gain knowledge about predatory patterns and procedures. As such, suitable semantic analysis methods facilitate discovering these grooming patterns. Significantly when changing the contexts of the

chat terms can change the whole meaning of a message. The meaning of the message is shown in the sentence or phrase, and the context is defined in a sentence manner rather than a single term. Therefore, this thesis proposes a model that extracts the embeddings for each sentence based on a Simple Contrastive Sentence Embedding framework (SimCSE) and uses the embeddings for training the machine learning techniques. The input sentences for each conversation can be encoded based on a pre-trained language model such as BERT or an improved BERT model called RoBERTa. The pre-trained networks based on BERT or RoBERTa can cope with cross-domain transferability for cyber grooming detection as it has very high bidirectional contextualization. Moreover, it allows the model to gain information from various representations with different positions in different contexts. The proposed model has gained an accuracy of 0.99% and an $F_{0.5}$-score of 0.99, which is the benchmark in predatory conversation detection.

Chapter 5

Limitations and Future Work

This thesis has examined several methods for online grooming detection in chat logs. We have proposed various approaches for authors profiling, predatory conversation detection, and predatory identification, along with some techniques to cope with the limitations in grooming datasets. This section provides details on the constraints that limit online grooming detection in real-life scenarios to give readers a deep knowledge of the problem in future work, considering their conditions and open research gap.

5.1 Cross-language Challenges

The distribution variation in feature vectors where the language of the data varies can cause a performance drop in machine learning models. Language syntax and semantics depend on the language family, and the approaches learned using one language may fail to scale up for another. It is problematic for grooming detection models to be trained on a different language from the testing dataset, where language variations, vernaculars, dialects, and country status can represent the conversation data differently. This issue arises more when most of the training models are based on English training models and are used to test another language, such as Norwegian. One can consider transfer learning to cope with cross-language text classification, where it does not require training and testing sets to be identically distributed.

5.2 Cross-cultural Challenges

The proper semantic analysis emphasizes the importance of cultural understanding in the theme of a language. It is a fact that the meaning and context of the same sentence can vary based on the culture of users. However, there is a lack of research

on how cross-cultural differences should be handled operationally for the semantic analysis of an online conversation.

5.3 Limited Understanding of Psychological Aspects

Gaining a profound knowledge of predators' modus operand and their motivation has the potential to improve detecting offenders before the crime happens. It is wise to perform an interdisciplinary approach in various areas such as psychology, linguistics, computer scientists, and law enforcement agencies such as the police to gain a better and more reliable model where we exploit complementary knowledge from different domains.

5.4 Deceptive Features

Online child molesters might abuse the knowledge about the children's patterns in writing. They can learn children writing styles and try to use this knowledge to conceal themselves behind imitated child patterns to avoid possible detections. The extracted features from the imitated behaviour are deceptive and challenge the performance and reliability of automatic grooming detection.

5.5 Generalizability of Grooming Detection Models

The generalizable grooming detection model requires the model to be tested on various types of datasets from different resources. It is challenging to test the generalizability of a grooming detection model due to the security and privacy issues of accessing predatory datasets.

5.6 Fusion

It is critical to collect the keystroke dynamics features along with the texts from online platforms. Chatlogs have both text features and keystroke dynamics characteristics. Applying both feature sets for training machine learning models increases the performance since they provide more information about the conversations. Therefore, it is wise to build a model that considers both feature sets for deciding if the chat is suspicious.

Part II

Published Articles

Chapter 6

Article 1: Online Grooming Detection: A Comprehensive Survey of Child Exploitation in Chat Logs

Borj, P. R., Raja, K., & Bours, P. (2022). Online grooming detection: A comprehensive survey of child exploitation in chat logs. Knowledge-Based Systems, 110039.

6.1 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.

6.2 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 behaviours by online extremism organizations [107, 108], disseminating fake news [109, 110] and spammers [111] to drastically impacting users' mental health [112]. Targeted chats can be used to spread hatred [113], manipulate the victim for propaganda, coordinate criminal or terrorist activities [107, 108, 3], radicalization, and in the worst of scenarios to target under-aged online users (minors and children) for sexual favors and abuse [4, 114]. 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 [107, 108, 115]. 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 [116].

Research on the retrospective view of online grooming experienced by minors showed that 25% of the minor participants talked with adult strangers [5, 6]. 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 [5]. 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 [6]. 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 [6]. 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 behaviours to lower the child's inhibition. Much research in

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

digital forensics has been produced in detecting online sexual predators. However, the majority of them have focused on children's images and videos [4, 10, 11, 7]. For instance, Lee et al. [7] 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 [7].

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 [117, 118]. 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 [12, 119, 120, 16]. 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 behaviour. 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 [121]. 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 [122, 123, 31]. For instance, Razi et al. [31] reviewed various approaches for sexual risk detection from a human-centered view considering sex trafficking, sexual harassment, and sexual grooming, and Miljana et al. [122] 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.

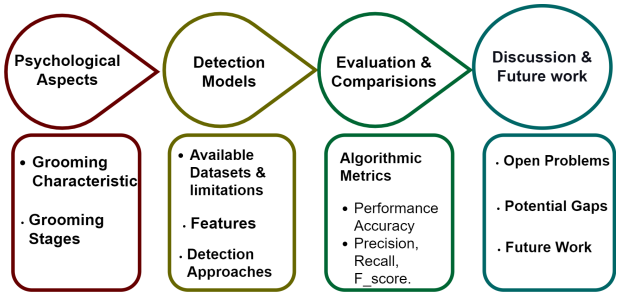

Figure 6.1: Overall Contributions of this Research Work

6.2.1 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.

Figure 6.1 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 [4, 10, 11, 7]. The taxonomy of this survey paper is illustrated in Figure 6.2.

The remainder of this paper is organized as follows. Section 6.3 first gives an overview of online grooming definitions and analyzes the psychological perspectives of sexual predators and different grooming characteristics. Section 6.4 breaks down the problem into different categories where Section 6.4.1 presents the available datasets for the research, and Section 6.4.2 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 6.4.3. The paper continues by surveying various grooming detection techniques in Section 6.4.4. Section 6.4.4 also describes how the grooming characteristics were used for detecting online grooming stages. Then in the same section, we detail previous research works for predatory conversation detection and predatory identification. Section 6.4.4 also summarizes the authorship profiling techniques for cyber-grooming purposes in online platforms. Section 6.5 discusses the challenges and open gaps of grooming detection and possible solutions along with its constraints, and finally, we conclude the paper in Section 6.6.

6.3 Online Grooming

6.3.1 Definition of Online Grooming

Online grooming is defined as a process performed by malicious actors such as pedophiles to entrap their victims [16, 17]. However, it is recommended not to use the term 'pedophile' to define grooming as it is used only after a precise clinical diagnosis and can not be used for all offenders [15]. One of the first comprehensive definitions of grooming was given by Craven et al. [15] 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*

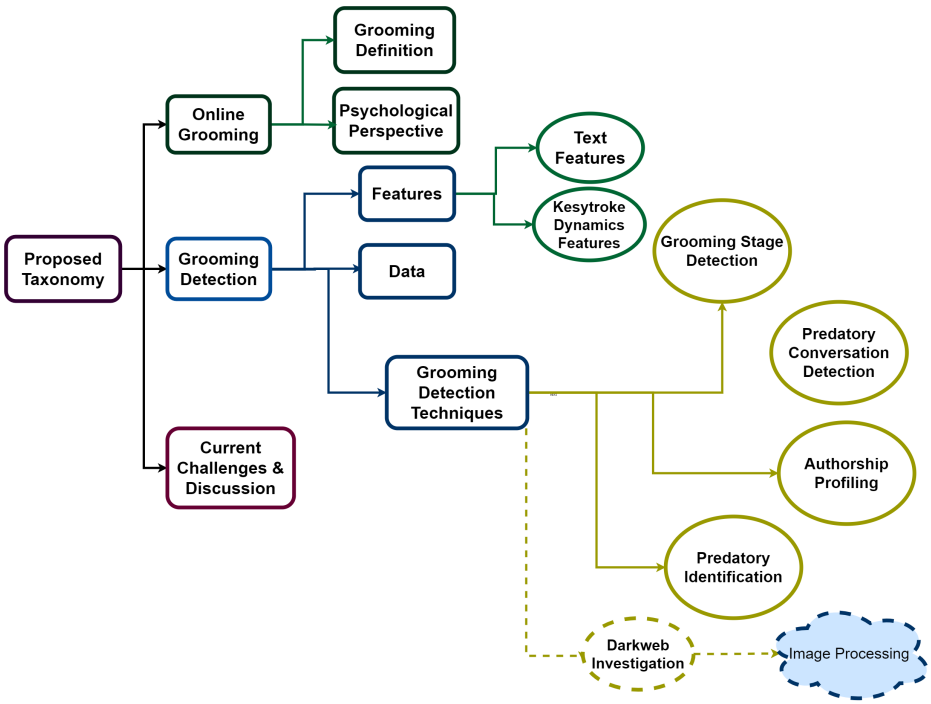

Figure 6.2: The proposed taxonomy for online grooming detection problem

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.

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 [36, 19, 37, 38]. 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.

6.3.2 Psychological Perspectives of Online Grooming

Child grooming has been researched in many works, including social and psychological areas [116, 124, 35, 125]. 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 com-

plex 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 [34]. In some cases of child abuse, it is challenging to detect the incident before the predator gains access to the victim physically [34]. 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 [124].

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) [125]. 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 [33]. Although they might have some similarities in conversing about traveling, parents, analyzing parents' work time, and sexual conversations about past relationships [85]. Whittle et al. [116] 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 [19]. 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 [19] 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 [19].

6.4 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 [38, 41, 30, 17, 38, 41, 42, 40, 71, 91, 65, 92, 24]. 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 [38, 41, 30] tried to detect online grooming with different perspectives. Some works [17, 38, 41, 42, 40] have explored various phases of grooming to investigate the different themes in online conversations, while others focused on predatory conversation detection [26, 25, 28], predatory identification [55, 50, 30], or author profiling [71, 91, 65, 92, 24].

6.4.1 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 [38, 41, 30, 17, 38, 41, 42, 40, 71, 91, 65, 92, 24, 26, 25, 28]. They have used different datasets for grooming detection, and Table 6.1 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.

Data	Sources	Ref
Predatory Data	www.perverted-justice.com	[16, 17, 38, 41, 30, 42, 40, 43, 44, 45, 46, 18, 47, 48]
	PAN2012	[49, 28, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63]
	MovieStarPlanet	[64]
Non-Predatory Data	www.literotika.com	[41, 42, 44]
	http://www.irclog.org	[30]
	http://krijnhoetmer.nl/irc-logs	[30]
	Omegle	[30]
	Twitter	[65, 66, 67, 68, 69, 70, 14]
	Blogs, Book Reviews	[18, 47, 71]
	British National Corpus(BNC)	[72, 73]

Table 6.1: A summary of datasets used for grooming detection in previous works

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. [18] 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. [44] used Literotica (www.literotika.com) data that contains conversations between adults that express their passion legally about sexual topics.

Recently, multiple works [28, 55, 50, 53, 52, 59] have used the PAN2012 [30] 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 [30]). 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 commu-

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

nication. 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. [64] used MovieStarPlanet as a source for data to detect grooming behaviour in online game platforms.

Datasets for Attributes Detection

Previous works [72, 73] 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. [72] 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 [126], and Lin [127] used the data collected from chat rooms for age detection and author profiling. Also, some previous research works [65, 66, 67, 68, 69] have used tweets to determine the online authors' demographic attributes.

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 [20], 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 behaviour 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 [27, 51, 62, 63].

- **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

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

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.

6.4.2 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 [28, 24, 53, 52, 51, 62, 63, 61]. 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, 10000 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) [128, 129, 130, 131]. BoW features for sexual predatory detection has been well used by a number of previous works [16, 51, 62, 63, 54, 50, 47, 61, 55, 26, 28, 53, 52, 48, 56, 64, 57, 60, 59].

Suppose that the data set D is defined as $D = X * Y$ where X represents the set of n documents, i.e. $X = \{d_1, d_2, ..., d_n\}$. Each document can be represented by an m-dimensional feature vector, where m is the size of the dictionary. Document i is represented as $d_i = (f_1^i, f_2^i, ..., f_m^i)$, where f_j^i (for $j \in \{1, ..., m\}$) represents the feature value for the j^{th} 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, 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 [28, 55, 50, 51].
- **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. [28] 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.

$$TF(t) = \frac{n_t}{n_d}, \tag{6.1}$$

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 [28, 55, 54].

– **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 [132]. 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 [132].

  TF-IDF computes the feature values based on the term frequency and the inverse of the document frequency [132]. Equation 6.2 shows how TF-IDF is computed:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{6.2}$$

  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:

$$w_{live,d_1} = 0.1 \times \log\left(\frac{3}{1}\right) = 0.158 \tag{6.3}$$

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 [62, 63, 54, 50, 61, 55, 26, 28, 52, 53, 60, 59].

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 [56, 57, 76]. Although Pendar [16] 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 [133] and GloVe [134] 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 [133]. Continuous Bag of word (CBoW) is the simple extension of a bigram model. Figure 6.3 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 [133]. For instance, suppose we have only one word w_t that

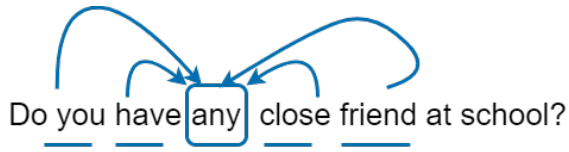

Figure 6.3: CBoW example

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 [133].

– **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

$$X_i = \sum_k X_{ik} \tag{6.4}$$

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

$$P_{ij} = P(j|i) = \frac{X_{ij}}{Xi} \tag{6.5}$$

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 [134]. So, the model for GloVe is extracted from the general model below (more details can be found in [134]):

$$F(\omega_i, \omega_j, \omega_k) = \frac{P_{ik}}{P_{jk}} \tag{6.6}$$

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 [28, 53, 58]. There exist other word embedding systems like BeRT [135], ELMo [136] and fastText [137], 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 [74]. 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 [74]. Capturing the psychological characteristics by the words can reveal the affective features of a conversation. Linguistic Inquiry and Word Count (LIWC) [75] 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.[60] 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 [26]. 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 [76]. The chat-based characteristics determine the type of conversation, for example, if it is negative or anxious. Parapar et al. [60] 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 behavioural 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 [138, 78]. 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 [138, 78] for an in-depth introduction to KD). Besides its natural use for authentication, KD can also be used to determine the authors' emotional state [139, 140, 141, 142], and emotional states can again help to detect the predators as they are not emotionally stable [76]. Borj and Bours [25] used KD feature vectors to detect authors that lied about their demographic information, such as age and gender, in chat conversations.

Figure 6.4 displays a summary of the proposed data and features.

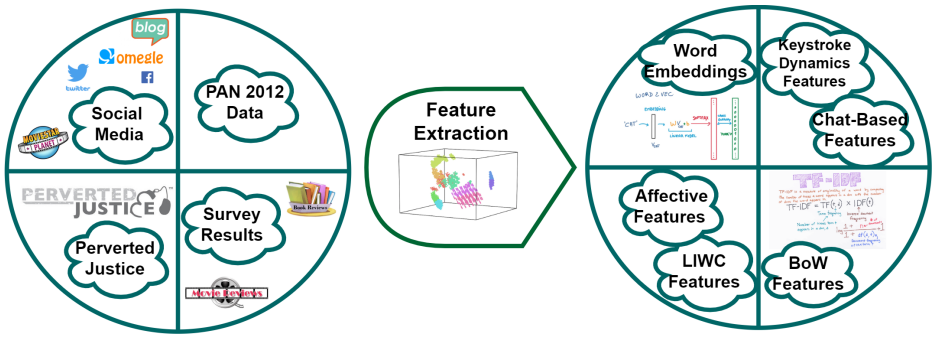

Figure 6.4: Data and Features in Sexual Predatory Detection

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 [51] 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), chi-squared, or frequency-based feature selection. For instance, Cardei and Rebedea [51] used MI to consider the power of presence/absence of a term in making the right classification decision for each class by using SVM [51].

Zuo et al. [62] 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. [62] 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 [62] 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) [62]. 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 [53]. 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 [53].

6.4.3 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.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{6.7}$$

- **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.,

$$P = \frac{TP}{TP + FP} \tag{6.8}$$

  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.

$$R = \frac{TP}{TP + FN} \tag{6.9}$$

  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

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \tag{6.10}$$

  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 6.2. The ranking for the PAN2012 competition was based on the $F_{0.5}$ score, so we kept that ranking here too. Further, Table 6.3 presents various works and their best performance for the problem based on the metrics mentioned above.

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

Participants	**Precision**	**Recall**	$F_{0.5}$	F_1	F_2
Villatoro et al. [61]	0.98	0.79	0.93	0.87	0.82
Snider *	0.98	0.72	0.92	0.83	0.76
Parapar et al. [143]	0.94	0.67	0.87	0.78	0.71
Morris & Hirst [57]	0.97	0.61	0.87	0.75	0.66
Eriksson & Karlgren [144]	0.86	0.89	0.86	0.87	0.89
Peersman et al. [71]	0.89	0.59	0.81	0.72	0.64
Grozea & Popescu *	0.76	0.64	0.73	0.70	0.66
Sitarz *	0.73	0.63	0.71	0.67	0.64
Vartapetiance & Gillam *	0.62	0.39	0.55	0.48	0.42
Kontostathis et al. *	0.36	0.67	0.39	0.47	0.57

Table 6.2: The best results reported on PAN2012 dataset, ranked based on $F_{0.5}$ scores. Note - * indicates no corresponding article available.

Ref	**Year**	**Accuracy**	$F_{0.1}$	$F_{0.5}$
Pendar et al. [16]	2007	-	0.94	-
Parapar et al. [143]	2012	-	0.84	-
Villatoro et al. [61]	2012	0.92	0.87	0.93
Bogdanova et al. [76]	2012	0.97	-	-
Cheong et al. [64]	2015	0.93	0.78	0.86
Ashcroft et al. [18]	2015	0.99	-	-
Ebrahimi et al. [52]	2016	0.99	0.77	-
Ebrahimi et al. [53]	2016	-	0.80	-
Cardei et al. [51]	2017	-	-	0.95
Escalante et al. [54]	2017	-	0.94	-
Zuo et al. [62]	2018	0.73	-	-
Zuo et al. [63]	2019	0.76	-	-
Misra et al. [56]	2019	-	0.58	-
Bours et al. [50]	2019	-	0.94	0.97
Borj & Bours [26]	2019	0.98	0.86	-
Muñoz et al. [58]	2020	0.88	0.42	-
Fauzi & Bours [55]	2020	0.95	0.90	0.93
Borj et al. [28]	2020	0.99	0.96	0.98
Ngejane et al. [59]	2021	0.98	0.70	-

Table 6.3: The accuracy obtained by various state-of-art works in detecting online grooming.

grooming detection is highly imbalanced, negatively impacting the accuracy metric's relevance [27]. 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 et al. [20], 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 [31].

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 [30] 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 [30] used F_β with $\beta = 0.5$ for ranking the performance results of the PAN2012 competition. Other researchers have followed this suggestion [61]. Some papers [50, 55] 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 [28, 26, 55, 50].

6.4.4 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.

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 [36, 39]. 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 [36, 39] 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 [36]. 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 [36, 39].

After finding the victim, the offender attempts to develop a trusted friendship. Olson et al. [36] 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 [19]. 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 [36, 19].

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 [37]. Predators also make the victims aware that their relationship is not appropriate to avoid jail in legal cases [37].

Generally, it can be said that sexual predators use different language themes in online conversations [93]. Each theme displays distinctive cognition used by online sexual offenders, such as discourse content, online solicitation, and fixated discourse [93, 145]. 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 behaviour in different patterns while minimizing the risk of being detected and preparing the victim mentally for further abuse [93, 145]. To understand trust in online predatory conversations, researchers found several patterns of compliments behaviour that show how by giving compliments, the predators frame the grooming process and gain the victims' trust [145].

Since online grooming has different stages, many researchers [38, 41, 42, 17, 40] 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 [38, 41, 42, 17, 40] have used is presented in Table 6.4.

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 Favourite 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

Table 6.4: Grooming stages and their characteristics

One of the first works for identifying the grooming stages was done by Kontostathis [38]. His tool annotated each message of the conversations based on the theory from Olsen et al. [36]. 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 [36]. To test the three stages hypothesis, Kontostathis et al. [45] 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. [45] 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 [38]. 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 [45].

Michalopoulos and Mavridis [46] 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 6.4.2) 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 [146] 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 [46].

Escalante et al. [49] 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 [49].

Online grooming was considered a four-stage process in [40]. 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) [147], 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) exclus-

ivity, (5) sexual, and (6) conclusion [17, 41]. Each stage contains some grooming characteristics presented in Table 6.4. For instance, in the friendship-forming stage, the features such as exchanging personal information like name, age, and location are considered [41].

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. [85] 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 [85].

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 [26]. 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 [41, 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?'* [41, 42].

The combinations of the words that indicate the various grooming stages are used as feature vectors to classify grooming stages in some papers [41, 42]. Gunawan [41] 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.

Gupta et al. [17] 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 [17]. 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 [17].

Figure 6.5 summarizes and shows the different stages of online grooming, and Table 6.5 gives a short overview of the most relevant papers.

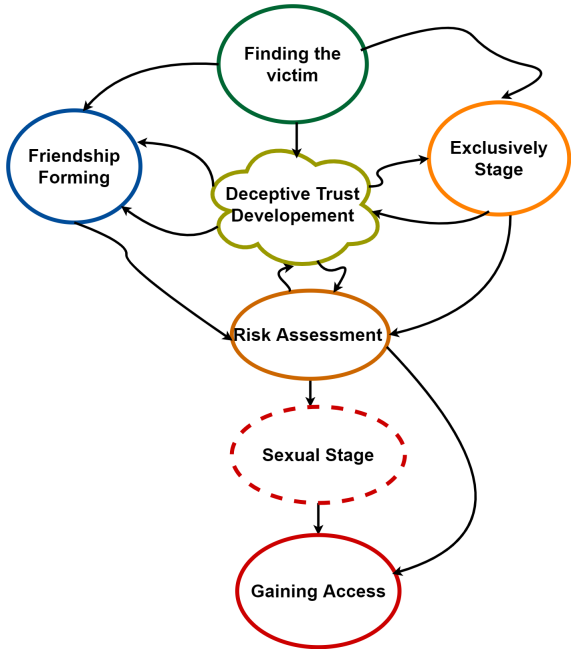

Figure 6.5: Grooming Stages

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. [61] 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 [61]. Fauzi and Bours [55] 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

Author	Number of Stages	Data	Features	ML Technique
Kontostathis [38]	3	www.perverted-justice.com	Groomng Characteristics	K-mean Decision Tree
Kontostathis et al. [45]	3	www.perverted-justice.com	Grooming Characteristics	Phrase matching Rule-based techniques
Escalante et al. [49]	3	PAN2012	BoW	Ring-based Classifier Decision Tree
Michalopoulos et al. [46]	3	www.perverted-justice.com	TF-IDF	Naive Bayes SVM Maximum Entropy EM and EMSIMPLE Decision Tree
Cano et al. [40]	4	www.perverted-justice.com	BoW, POS, sentiment, complexity, readability, length, psycho-linguistic (LIWC dimensions)	Decision Tree SVM
Gunawan et al. [41]	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. [17]	6	www.perverted-justice.com	LIWC to create psycho-linguistic profile based on grooming characteristics	Logistic Regression Decision Tree

Table 6.5: Online Grooming Stage Detection by different papers

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 [55].

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. [28] 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, 28]. The authors showed how preprocessing could increase the performance of the models for sexual predatory conversation detection [28].

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 [148]. Cardei and Rebedea [51] tested several techniques to cope with the imbalanced data in grooming detection problem: cost-sensitive technique, sampling techniques such as BalanceCascade [104], and clustering-based method using k-means [105]. 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. [63] 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 [54, 50, 86]. 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. [86] 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. [54] 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 [50] 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 [50].

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 [47]. 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 [47]. 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 [149]. Java WordNet Similarity library (a Java implementation of Perl Wordnet) [150] and Resnik's similarity measure [151] are used for extracting affective features [47]. Applying the affective feature sets showed that it is challenging 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 [47].

Some chat conversations may implicitly show signs of predatory behaviour 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. [58] 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 [58]. 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. [52] 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. [53] used CNN for predatory chat detection. The CNN model gained a better performance compared to the semi-supervised anomaly model.

Predatory Identification

Few earlier works mentioned above have also tried to identify the predators [61, 50, 28, 55]. From a forensic's perspective, it is crucial to identify the predators for further actions. This section will discuss various methods [16, 64, 57, 88, 60, 18] that have been used for predatory identification.

Pendar [16] proposed one of the first models for distinguishing online predators from victims, using SVM and k-Nearest Neighbors (KNN) machine learning methods. He [16] used the Perverted Justice website to collect the data and extracted n-gram features, including document frequency and character ratios. Similarly, Cheong et al. [64] tried to detect the predators in online game chat platforms, applying a combination of inherent features with the BoW representations. Morris [57] 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 behavioural features that cap-

[4]The details of the implementation by Munoz et al. [58] can be found on the GitHub repository https://github.com/gisazae/Tensorflow-Examples/blob/master/IntegracionCorpus_checkpoint.ipynb.

ture the author's conversation flow pattern, such as turn-taking and message length [57].

Objective		
Grooming Stage Detection	**Predatory Conversation Detection**	**Sexual Predatory Identification**
Cano et al. [40]	Bogdanova et al. [47]	Ashcroft et al. [18]
Egan et al. [93]	Borj et al. [28]	Borj et al. [28]
Escalante et al. [49]	Bours & Kulsrud [50]	Cardei & Rebedea [51]
Gunawan et al. [41]	Cardei & Rebedea [51]	Cheong et al. [64]
Gupta et al. [17]	Ebrahimi et al. [53]	Fauzi & Bours [55]
Kontostathis [38]	Ebrahimi et al. [52]	Misra et al. [56]
Kontostathis et al. [45]	Escalante et al. [54]	Morris [57]
Michalopoulos & Mavridis [46]	Fauzi& Bours [55]	Pendar [16]
Pranoto et al. [42]	Miah et al. [48]	Villatoro et al. [61]
	Misra et al. [56]	
	Munoz et al. [58]	
	Zuo et al. [62]	
	Zuo et al. [63]	

Table 6.6: Categorization of works according to objectives in predator detection.

From a psycholinguistic perspective, there is a relationship between word usage and personality, social conditions, and consequently emotional states [88]. 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. [60] extracted various features exploring this concept to distinguish predator behaviour from a victim's behaviour. 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 [60]. 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.[60] 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. [18] 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 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 [18].

Table 6.7 presents all the related works for grooming detection with a focus on predatory conversation detection and predator identification. Finally, table 6.6 presents a taxonomy classification of the proposed approaches for grooming stage detection, predatory conversation detection, and predatory identification.

Author	Objective of the research	Data	Features	ML Technique
Ashcroft et al. [18]	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. [47]	Predatory Conversation Detection	NPS chat corpus, Cybersex Logs, http://perverted-justice.com	Emotional Characteristics, n-grams	SVM
Borj et al. [28]	Predatory Conversation Detection & Predatory Identification	PAN2012	BoW, TF, TF-IDF, GloVe	SVM, NB, RF
Bours et al. [50]	Predator Detection	PAN2012	BoW, TF-IDF	LR, Ridge, NB, SVM, NN
Cardei et al. [51]	Predatory conversation detection & predator identification	PAN2012	BoW & behavioural characteristics	SVM
Cheong et al. [64]	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. [52]	Predatory Conversation Detection as an anomaly	PAN2012	TF-IDF	1-class SVM
Ebrahimi et al. [53]	Predatory Conversation Detection	PAN2012	GloVe, Word2Vec, BoW	SVM, CNN, NN
Escalante et al. [54]	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. [55]	Predatory Conversation Detection & Predatory Identification	PAN2012	BoW, TF, TF-IDF	NB, SVM, NN, KNN, RF, Ensemble Model
Miah et al. [48]	Predatory Conversation Detection	http://www.fugly.com, http://chatdump.com, http://perverted-justice.com	Term-based features, Psychometric Characteristics	NB, Regression
Misra et al. [56]	Authorship Attribution of Online Predatory Conversations	PAN2012	Character unigram and bigrams	CNN
Morris et al. [57]	Predatory Identification	PAN2012	Lexical features, Behavioral features	SVM
Muñoz et al. [58]	Grooming Detection	PAN2012	Word2Vec	CNN
Ngejane et al. [59]	Predatory Conversation Detection	PAN2012	TF-IDF, Embedding	LR, XGBoost, MLP, BiLSTM
Parapar et al. [60]	Predatory Identification	PAN2012	TF-IDF, LIWC, chat-based & Content-based features	SVM
Pendar et al. [16]	Predatory Identification	http://perverted-justice.com	n-grams	SVM, KNN
Villatoro-Tello et al. [61]	Predatory Identification	PAN2012	BoW, TF-IDF	NN,SVM
Zuo et al. [62]	Grooming Detection	PAN2013	BoW, TF-IDF	GNB,LR, AdaBoost, Fuzzy Interpolation
Zuo et al. [63]	Predatory Conversation Detection	PAN2013	BoW, TF-IDF	ANN

Table 6.7: Summary of research works on Sexual Predatory Conversation Detection and Sexual Predatory Identification

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 [25, 24]. 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 Figure 6.6).

It is logical to consider stylometry for author profiling, given that predator detection in online conversation is conducted using chat logs [72, 73]. 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 [81, 89, 90, 84, 82, 83]. Figure 6.4.4 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.

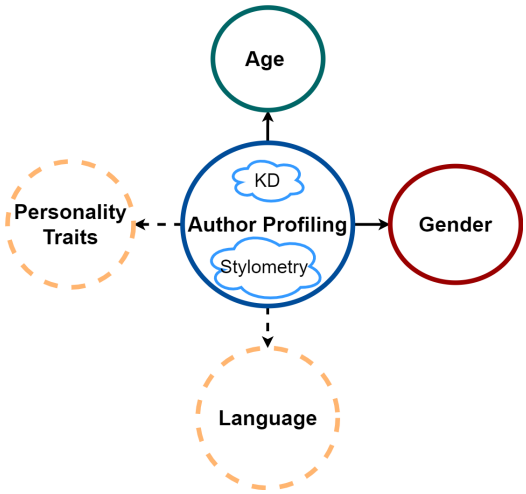

Figure 6.6: Author Profiling

- **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 [81, 89, 90]. Giot

and Rosenberger [81] 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 [89] 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 [89]. Idrus et al. [90] 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 [84, 25, 82, 83]. Pentel [82] 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. [84] 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. [83], 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. [72] and Argamon et al. [73] respectively. Koppel et al. [72] detected the gender of the authors on the BNC[5] corpus by extracting simple lexical and syntactic features with an accuracy of 80%. Argamon [73] 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 [71, 91, 65, 92]. Peersman et al. [71] provided a method to distinguish between adults

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

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 [91]. Nguyen et al. [65] 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 [65]. They also considered the impact of gender on the performance of age detection, where age and gender are assumed as inter-dependent variables, and the obtained correlations were around 0.74 [65]. PAN[6] held a series of competitions for author profiling over recent years. In PAN2013 [92], 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 [152], 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 [70] and 2017 [14] was also extracted from Twitter. The purpose of PAN2018 [153] was to detect the gender by texts and images. The data was based on the PAN2017 [14] 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 [92, 66]. A combination of content-based and style-based features provides a good clue for gender and age detection [66]. 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

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

on the data and research goal [66, 68]. 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 [154]. Peersman et al. [71] used n-grams of words or characters and morphological, lexical, and semantic features for recognizing adults versus adolescents.

The participants of the PAN2013 competition [92] 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 [92]. The PAN2015 participants [152] 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 [155]), or LIWC [152, 156, 157]. The goal of the PAN2016 [70] 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 [70].

Basile et al. [158] 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 [158]. 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 [158]. PAN2018 [159], 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 [159], while some represented the documents with word embedding feature space [160, 161]. In [162], the authors combined the POS tag n-grams with syntactic dependencies for gender detection to capture the verbal constructions. Daneshvar and Inkpen [163] 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. [164] to increase the relevant groups' security by highlighting the target group in the early stages [164]. 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 [164].

– **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. [24], 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 [24].

Table 6.8 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:

  – **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

Authorship Profiling Method	Author	Object	ML Technique
Keystroke Dynamics	Giot et al. [81]	Gender Detection	SVM
	Fairhurst et al. [89]	Gender Detection	K-NN, DT, NB, Fusion Models
	Pentel et al. [82]	Age & Gender Detection	SVM, K-NN, RF
	Tsimperidis et al. [84]	Age Detection	ANN
	Tsimperidis et al. [83]	Age Detection	RF, SVM, NB, Multi-Layer Perceptron, RBF
Text Analysis	Argamon et al. [66]	Age & Gender Detection	Multinomial Regression (BMR)
	Koppel et al. [72]	Gender Detection	Exponential Gradient Algorithm
	Schler et al. [68]	Age & Gender Detection	Multi-Class Real Winnow (MCRW)
	Goswami et al. [154]	Age & Gender Detection	Naive Bayes
	Nguyen et al. [65]	Age Detection	Logistic & Linear Regression
	López-Monroy et al. [164]	Early Author Profiling	SVM, Naive Bayes
	Basile et al. [158]	Gender & Language Detection	SVM
	Peersman et al. [71]	Age & Gender Detection	SVM
	Daneshvar et al. [163]	Gender Detection	SVM
Both	Li et al. [24]	Gender Detection	Random Forest

Table 6.8: A summary of Authorship Profiling Papers

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 [165]. Note that the data size and domain similarities substantially impact the performance of gender detection in cross-domain gender detection [165].

– **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.

6.5 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.

6.5.1 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 [16, 30].

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 [20, 51, 62]. 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 [27]. 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 [103]. 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 [27].

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 [166, 167]. 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.

6.5.2 Topic and Context Modelling

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 [93]. 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 [19]. To minimize risk, predators express their emotions so that the victims trust them while their primary incentive is hidden [145]. 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 [168]. 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.

6.5.3 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 [169]. 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 [106].

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) [135, 168] 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 [106].

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

6.5.4 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 [170, 171, 172, 173]. 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 [174]. 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 [170, 171, 172, 173]. 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 [175].

6.5.5 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 [176]. 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 [177]. They might not show their actual behaviour due to lack of trust, social desirability, and fear of losing anonymity [177]. 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 [176].

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.

6.5.6 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. [17] 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. [121], 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 [178, 179, 180]. For instance, Michalopoulos et al. [180] 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. [179] 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.

6.5.7 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 behaviour 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 [181, 182, 25]. 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 behaviour by the groomer is inevitable, and implementing an author profile model that detects deception remains challenging.

6.6 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 under-

standing of predatory behaviour. 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.

Chapter 7

Article 2: Detecting liars in chats using keystroke dynamics

Borj, Parisa Rezaee, and Patrick Bours. "Detecting liars in chats using keystroke dynamics." Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications. 2019.

7.1 Abstract

In this paper we will investigate the possibilities for detecting liars in chat rooms who have taken on a different identity. While using a different identity people might require more time to reply to questions of the chat partner, or might use corrections to change their text to avoid inconsistencies in their answers. These issues will cause differences in the typing behavior, which can be measured in the typing rhythm. We have shown in this paper that, with a high accuracy, we can distinguish between a chat of a person who uses his/her own identity and is honest in his/her answers, and a chat of a person who is lying because his/her answers need to be consistent to an assumed identity. We obtained a correct classification of a single message in a chat with an accuracy of more than 70% and a correct classification of a full chat with well over 90% accuracy.

7.2 Introduction

The Internet is used widely by children as a form of entertainment rather than an environment to research. They use it to communicate or play games with their friends. It can be said that there is no difference between online and offline worlds for children. Social networks are parts of their lives while they are experiencing something new every day. Online a person can be wherever and whatever he or she

wants to be. However, anonymity carries many risks for children since pedophiles can easily find their victims on the internet while they have encloaked themselves with fake identities. It is easy to access the internet through smartphones, and it increases the amount of online grooming. According to reports in 2018, the average time children spend online per week is 20 hours for the children between the age of 12-15 and more than 13 hours for the kids between the age of 8-11. It also shows that 24% of young people have experienced an adult that they don't know in real life trying to contact them online, and 7% of the children under 13 have been asked for sexual images or messages [183].

It should be noted that children not only can be sexually or physically abused but also emotional abuse or neglect can have a significant impact on their lives [183]. As an example of this occurrence, we can consider the MySpace case where Megan Meyer started chatting with a teenage boy named 'Josh'. After a while, Josh abruptly ended their relationship by telling that the world would be a better place without her, which led to Megan committing suicide. Police discovered that Josh was a 52 years old woman who pretended to be a teenage boy!

People can lie about their real identities such as age and gender, and the question is how can we detect such lies in a chat room? It is evident that the conventional ways for lie detection such as polygraph are not useful in social networks and chat rooms. According to Walczyk [184], the polygraph is a device that continuously records psycho-physiological arousal as assessed by pulse rate, blood pressure, respiration rate, and skin conductivity, which has been applied to uncover deception. Polygraph test includes two parts. In the first part, some questions are given to the examinee to get some basic information which can be used for controlled questions. During the second part, the examination part, some relevant questions will be asked besides the control questions, where people might give false answers. The examinee is finally given the question that is related to the issue under investigation [184].

Lie indicators are typically faint and unreliable in online communities, and so, it is difficult to detect the deception in chat rooms. Psychological works have shown that fraud is more cognitively demanding as the liar must provide a new story and memorize everything not to make any mistake, and consequently, he or she needs more time to think and process the conversation [185]. So, capturing the time during the conversation in a chat room can give us some indication if someone is a liar.

The primary goal of this paper is to provide a model can detect liars in chat rooms, and we want to investigate if typing behavior can support the cognitive lie detection perspective by capturing time features during an online conversation. In

biometrics, Keystroke Dynamics is a technique to authenticate or identify a user based on his/her typing behavior. In this paper, we will use Keystroke Dynamics as the term for typing behavior for lie detection too.

Contrary to the intrusive nature of the common methods for lie detection, Keystroke Dynamics can be collected even without the knowledge of the user [186]. Keystroke Dynamics has also been used widely for other goals, for example, it was used to detect the emotional states of the users based on the rhythm of typing on the keyboard [142, 139, 140, 141], and various tools using Keystroke Dynamics were proposed to detect the gender [89, 81, 83]. Additionally, Pentel [82] and Tsimperidis [84] tried to identify the age of the user by Keystroke Dynamics.

The remainder of this paper is structured as follows: Section 7.3 overviews related work on the various methods of deception detection, and Section 7.4 outlines the methodology including the background hypothesis and experimental setup. Our approach to the problem of lie detection in online conversations is presented in Section 7.5. We show the results of the conducted experiments in Section 7.6. Finally, we present a discussion on the research, as well as conclusions and suggestions for future research in Section 7.7.

7.3 State of the Art

Different types of neurological and physiological indicators have been introduced for deception detection such as skin conductance, heart rate, respiration, pulse volume, facial temperature, etc. However, these type of indicators have some restrictions. For example, some attributes of physiological and neurological signs such as habituation and non-responding can spoil the reliability of the tests based on these indicators. Also, to use these indicators for lie detection, expensive types of equipment are needed to capture the physiological signals [187]. Therefore, research interest in using behavioral indicators for lie detection has significantly increased. Response latency or Reaction Time (RT) is one of the behavioral methods, which recently, has been studied widely [101].

Cognitive theories assume that people need more time while they are lying as lie production requires more cognitive effort because of the complexity, stress, and threat of detection [98, 99, 100]. So, lying increases the memory load, and consequently, increases the reaction time [97]. Although, this assumption is not always valid as some studies showed that not only lying becomes more difficult if people are trained to tell the truth but also that lying becomes easier if individuals are trained to lie more often [188].

Keystroke Dynamics (KD) is a behavioral measurement which identifies or recognizes users based on the way they type on their keyboards. In KD features based

on the time of typing are extracted, and it can capture response time and latency while an individual types on a keyboard. Thus, some studies applied Keystroke Dynamics for lie detection. According to the model, which Grime [102] proposed to show the relationship between lying and Keystroke Dynamics, lying affects the emotional arousal and cognitive load, and these effects can change the behavior in typing on the keyboard [102, 97]. Similarly, to show if lying increases the cognitive load, Derrick [99] proved that deception is correlated with behavioral metrics. He analyzed the response time, the number of edits, and word count and lexical diversity which were captured by using Chatterbot.

To distinguish a truthful text from a deceptive one by KD, Banerjee [186] used data where people were asked to write honest and deceptive texts about one of these topics: the restaurant review, gay marriage, and gun control. Then, for illustrating the differences in editing and timing patterns in two different types of texts, they extracted some features such as the number of deletion keystrokes (Delete and Backspace), the time-span features such as duration and latency of various linguistic units (e.g., words, sentences, and entire documents), and differences in the speed of writing of various words such as nouns, verbs, adjectives. They classified the data in two truthful and deceptive texts using binary SVM with an accuracy of 83.62% [186].

Recently, Monaro [189] used KD for identity identification. They asked participants to answer the questions about their autobiographical information under two different conditions: true condition and false condition. In the false state, the authors provided a fake identity for each participant for answering the questions while in a true condition, participants had to give answers based on their real autobiographical information. The questions included some information such as identity, physical characteristics like gender and age, residence, contacts, etc. While participants were using fake identities, they were asked some unexpected question like *What is your zodiac sign?* or *Which is the capital town of your residence region?*, and apparently, answering these questions with a false identity required more time. Monaro extracted some features such as the time between the first key and ENTER, average typing speed, answer length, duration, and latency. Various types of machine learning methods such as Support Vector Machine (SVM), Random Forest (RF), Logistic Model Tree (LMT), and Logistic were used, and the accuracy of detecting liars from honest people was around 95% [189].

7.4 Method

7.4.1 Hypotheses

According to Derrick's results [99], we assume that there are some differences in the behavior of a person when lying or when telling the truth in a chat room. It was shown that deception increases the cognitive load, and therefore, it increases the time to answer and communicate in a chat room. So, the first hypothesis is that writing in a chat-based message on average takes more time while people try to lie.

The second hypothesis is that while users are lying, they edit the text more than they would when providing valid information. For example, users evaluate whatever they have written and try to make their lies more deceiving, and as a result, they might start editing and deleting whatever looks unreasonable, and in addition, because the text is reviewed more thoroughly, it will take longer time between finishing the typing of the message and sending it.

The last hypothesis is based on the fact that generating lies is more complicated than recalling valid information from memory. Furthermore, this difficulty decreases the ability of the liar to make complex sentences during a chat session, and then, it is assumed that the messages are shorter while the user is lying [99].

7.4.2 Experiment

We performed an experiment which captured KD information of users while they were chatting. The goal of the test is to reproduce as much as possible the reality of initial online contact between two people that do not know each other. We used Skype as a chat application and the KD information was collected using an application called BeLT. BeLT (Behavioural Logging Tool) has been developed at the Norwegian Information Security Laboratory (NISlab) in 2013 as part of a bachelor student project [80]. When a subject uses his or her computer, BeLT collects all the interactions with the keyboard and the mouse, as well as some information about software and hardware. In particular, all interactions are registered in a chronological list which displays a variety of information. As for the keyboard interactions, BeLT captures the key code (which key is pressed or released), the event (either KeyDown or KeyUp), and the timing value associated to when event takes place.

The series of experimental sessions took place at NTNU, where two separate rooms were equipped with hardware devices. The rooms were located such that the chatters could not see each other. In order to avoid possible device-specific variation, the chat experiment was performed using an identical set of tools such

as the same laptop, and the same version of Skype. To block all ads, a custom host file was used, and special care was taken to ensure that these specific restrictions were blocking all ads while not interfering in the conversation in any way. Therefore, every participant had the same screen to chat.

In the experiment the participants were asked to chat with a fixed moderator. The chat experiment included two components: a true component and a lie component. In the true component, the participants were asked to remain honest during the conversation with moderator. In order to avoid lying during this chat, the participants were instructed to inform the moderator if the did not want to reveal specific personal information or did not want to continue talking about a specific topic.

For the lie component, fake identities were provided for each participant, and the participants were asked to adopt the given fake identity. In particular was the fake identity always from a child (all the participants were adults) and the gender of the fake identity was the opposite of the gender of the participant. So, this component required participants to lie about their real identity and pretend to be another person, and required them to tell lies during the chat.

7.4.3 Participants

There were thirty-three participants, including 13 females and 20 males. Thirty-two of the participants are students of NTNU, and is an employee at NTNU. The average age was 30, and the average education was 18 years. Each participant signed an consent form which described the procedure, and goal of the experiment, and the participants were informed that their data was handled anonymous and that they could withdraw at any moment during and after the experiment. The participants also provided some demographic information such as age and gender. From the list of ages of the participants in Figure 7.1 we can see that the ages ranged from 22 to 44.

7.5 Data Analysis

7.5.1 Feature Extraction

To analyze the data, it is needed to build a set of features from the raw KD information, which can be used for training the model. Each conversation was split into a set of messages, and for every single message, the following two main timing features were extracted:

- Duration: This feature defines the time between each key is held down, i.e. from the time it is pressed down till it is released;

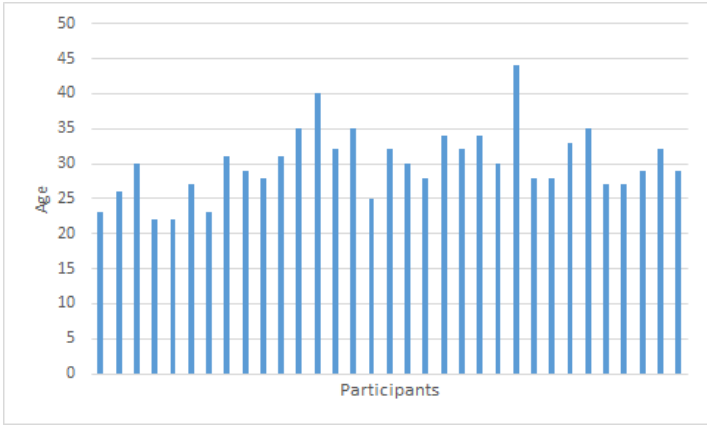

Figure 7.1: Age Distribution of the Participants

- Latency: This feature equals the time between two consecutive keys, i.e. from the first key is released till the next key is pressed down.

Besides the above defined latency there are 3 alternative definitions of latency that we also use:

- PP-latency: The latency when pressing down two consecutive keys;
- RR-latency: The latency between releasing two consecutive keys;
- PR-latency: The time between pressing down the first key and releasing the next key.

The above 5 timing features are illustrated in Figure 7.2

In addition to the features in the Figure 7.2, for every single message, the following features were extracted as well:

- Length feature: The length of the message, i.e. the total number of keys pressed, including special keys (e.g. space, shift, backspace, delete, etc.);
- Special keys features: The frequency of using special keys such as backspace, left shift, right shift, and space;
- Pause time: We compute the time of releasing the key before the space and pressing down the key after space and call it space-time. To clarify, space-time is the time that no other key than the space is typed. From this space-time feature, we extract two different features including short space-time

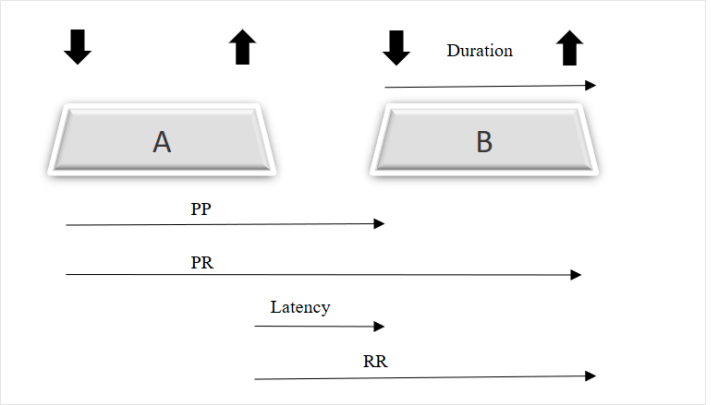

Figure 7.2: Keystroke Dynamics Timing Related Features

and long space-time, which indicates a thinking pause. It is assumed that if space-time is more than 150 ms, we consider it as a long-space time. When users are lying, they might pause longer and more often to think about what to write as a deceiving lie;

- Backspace features: There are two different types of features which are related to using the backspace. The first one shows the time it takes to consider using backspace (i.e. to determine to change the given text), and the other indicates the time the user needs to start typing again after the last used of backspace (i.e. to determine the remainder of the text);

- Shift features: The set of features which are related to using left shift and right shift: this set includes average time of duration and latency of shift keys, and relative use of right shift vs. left shift;

- Enter feature: This feature indicates the time between releasing the last text key of the message and pressing the enter key, i.e. the pause taken before sending the message;

- Tri-graphs features: In Figure 7.3 consider the middle key B and the time elapsed from when the previous key (A) is released, till the next key (C) is pressed.

7.5.2 Feature Selection

The data was split between training and testing data and based on the training data, all the features were scaled between 0 and 1. Scaling was done by a linear mapping

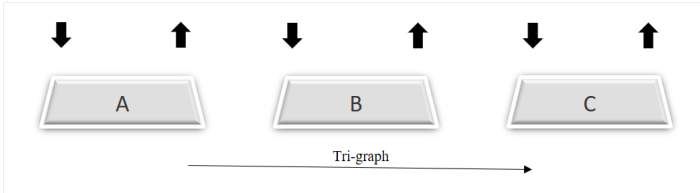

Figure 7.3: Tri-graph Feature

where the minimal value for a feature was mapped to 0 and the maximal value to 1.

It is evident that many features are dependent, and so, there are overlapping features. To remove the redundancy and decrease the dependency between features, PCA was run on the data to reduce the number of features. Analysis has been performed with and without PCA to see the difference in performance.

7.6 Results

We have analyzed our data in various manners, which will be described below.

7.6.1 Message-based Scenario

In this scenario, we split all the conversations of all participants into a set of messages, where every single message will be a sample. The before mentioned features were extracted for each message and PCA was applied to reduce the dimensionality of the feature set. Each message is tagged based on the conversation condition, i.e. as a lie or a truth. The goal of classification in this scenario is to recognize if a random sample is a lie or a truth. The data was split into two sets including training data and test data, where 70% of the data was used for training and the remaining 30% for testing.

Classification Result

We applied different machine learning algorithms, and Support Vector Machine (SVM) appears to provide good results among the selected methods of computation. In many papers on KD research, SVM was also used for binary classification such as gender classification and age classification. Besides SVM we also applied Decision Tree (DT) and Naive Bayes (NB), but performance of these methods did not give any significant improvement over SVM, as can be seen in Table 7.1.

To train the SVM, we used Radial Basis Function kernel (RBF), and 10-fold cross-validation on the training data. The results of our evaluation are summarized in Table 7.1 which includes accuracy, precision, recall, and F-score for each of the

three applied methods. Here accuracy refers to the fraction of correctly classified samples. Precision is the probability that a unknown sample that is classified as a lie is indeed a lie, while recall indicates the probability that a lie message will indeed be classified as a lie. Finally, F-score is defined as the harmonic mean of precision and recall and is calculated according to the following equation:

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

From Table 7.1 we find that SVM gave the best performance results for classification a single message with the accuracy of over 70%.

Method	Accuracy	Precision	Recall	F-score
SVM	0.72	0.74	0.58	0.65
NB	0.65	0.69	0.54	0.61
DT	0.66	0.64	0.63	0.63

Table 7.1: Performance results for the message-based scenario

Training size	MCV	SSC	LSSC(0.2)	LSSC(0.3)	LSSC(0.4)	LSSC(0.5)
24	0.76	0.89	0.84	0.83	0.77	0.74
25	0.71	0.93	0.84	0.84	0.91	0.84
26	0.69	0.91	0.90	0.93	0.89	0.78
27	0.94	0.84	0.91	0.88	0.83	0.81
28	0.81	0.88	0.80	0.80	0.82	0.96
29	0.98	1.00	0.90	0.90	0.92	0.88
30	0.81	1.00	0.97	1.00	1.00	1.00

Table 7.2: Performance accuracy result for the chat-based scenario

7.6.2 Chat-based Scenario

In this scenario will we also classify separate messages as a lie or a truth, but we will group them based on the conversations. We will use the classifications of each of the separate messages in a chat to determine if the participant was lying or telling the truth in that conversation, i.e. if the chat was part of the true or the lie component of the experiment. Given that in the previous scenario SVM performed best have we restricted this analysis to SVM only. The data of the 33 participant (33 lie chats and 33 truth chats) have been split into a training and testing set, where the size of the training set ranged from 24 to 30 participants. In particular, if the the number of participants for the training set is set to N, then N out of the 33 participants are selected randomly and both the lie and truth chat

of these participants are used for training the SVM model. The 2 chats of each of the remaining (33-N) participants are used for testing, i.e. we have M=2*(33-N) chats for testing.

When testing a single chat, all the messages in this chat were classified according to the trained SVM model. The chat itself was classified in different manners, based on the results of the classification of each of the separate messages in the chat. We have tested with the following methods:

1. **Majority Classification Voting (MCV):** In this case the majority of the classifications of the messages determined the classification of the chat. In other words if the majority of the messages in a chat were classified as a lie, then the chat was classified as a lie.

2. **Sum Score Classification (SSC):** The SVM algorithm returned a score between -1 (lie class) and +1 (truth class), and we calculated the sum of all the scores of the messages in a chat. If the total sum was negative, then the chat was classified as a lie, and with a positive sum the chat was classified as a truthful chat.

3. **Limited Sum Score Classification (LSSC):** The returned classification score from the SVM model gives an indication how "certain" the classification is. In case of a score close to -1, the message is highly likely a lie, while a negative score close to 0 is more doubtful. In the above 2 methods (especially the Majority voting), this fact is not regarded. In this Limited sum score we therefor ignore scores produced by the SVM model that are too close to 0 and only regard scores that are above a threshold δ in absolute value. In other words only classification scores above δ or below $-\delta$ are summed. As before, a negative sum will indicate a lie chat, while a positive score will indicate a truthful chat. We have tested this method with different values of δ.

Classification Result

We have run the analysis both with and without using PCA. First we present the results in Table 7.2 where PCA was applied to reduce the dimension of the feature vectors and to remove dependency between the various features that are used. The first column will indicate the number of participants that were randomly selected for training the SVM model (hence the number of chats used for training is twice as much). We see that the accuracy improves over the 72% mentioned in Table 7.1 for SVM. From Table 7.2 we can see that the performance of MCV is outperformed by SSC, which was to be expected from the argumentation given above. When

Training size	MCV	SSC	LSSC(0.2)	LSSC(0.3)	LSSC(0.4)	LSSC(0.5)
24	0.81	0.83	0.79	0.87	0.80	0.79
25	0.72	0.87	0.81	0.78	0.85	0.87
26	0.78	0.86	0.84	0.86	0.88	0.86
27	0.74	0.91	0.79	0.89	0.91	0.86
28	0.74	0.91	0.79	0.89	0.91	0.86
29	0.74	1.00	0.81	1.00	0.99	0.96
30	0.99	0.99	0.91	0.93	0.90	0.92

Table 7.3: Performance accuracy result for the chat-based scenario (no PCA)

the SSC method is compared to the LSSC method, for the 4 selected values of δ, we see that using the limited sum score does not really improve on the (unlimited) sum score classification.

The results in Table 7.3 are obtained in the same manner as above, except that no PCA has been applied to reduce the dimensionality of the feature set. We note that the same observations from above can be drawn from Table 7.3, i.e. LSSC does not improve over SSC and the sum score classification method gives higher performance accuracy than simple majority voting. Also here the accuracy of correctly classifying a chat is higher than the accuracy of correctly classifying a single message as shown in Table 7.1. When comparing the results of Tables 7.2 and 7.3 we can conclude that the accuracy is slightly higher when using PCA. We do need to remark that the number of test samples is rather low. Specifically when using $N = 30$ participants for training, only $M = 3$ participants are available for testing, i.e. the number of chats for testing is only 6, even though each message within these 6 chats is classified separately by the trained SVM model.

7.7 Conclusions and Future Work

' In this section we will first discuss the work we have done in this research, before giving conclusions on the results. We finally give suggestions where we can extend on this work.

7.7.1 Discussions

In this research we wanted to distinguish between honest and dishonest people in a chat environment, based on the typing dynamics. We conducted an experiment where a moderator chatted over Skype with 33 participants in two different sessions. In the first session the participants acted honestly, while in the second session they had to assume a different identity, where they were supposed to be a child with a gender different from the actual gender of the participant. During the

chat sessions, the moderator noticed that the reaction time often was slower when the participants had assumed the alternate identity. Also noticeable was the fact that it often was harder to get replies from participants in the dishonest session. This is not entirely surprising, because most people will not know how to think as a child of a different gender.

Even though a participant has assumed a different identity is not every message that is send a lie. General messages like greeting or saying goodbye are neutral, yet they are included in the analysis here. Questions that were hardest to lie to, where opinion-like questions (e.g. what movies a person liked) or derivative questions (e.g. in what year a person was born when the age was provided). For the first type of question, the participants had to come up with an answer that would fit the new personality, e.g. the favorite book of a 12 year old boy would most likely not be "Wuthering Heights" or "Pride and Prejudice". The second type of question would require the participant to think such that not a answer is given that contradicts previous information.

In our analysis we restricted ourselves to the KD information about the typing rhythm of a participant. Alternatively we could have looked at the actual texts of the messages and extract features from those. At a low level this could include features related to average word length, number of words in a sentence, and use of punctuation and emoticons. At a higher level understanding of the meaning of the texts could be included too in the analysis.

7.7.2 Conclusions

From the result, we can see that there are differences in the behavior between a person who lies and one who is honest in a chat room. This conclusion is based on typing behavior, such as making corrections, typing speed, and pause times. We have shown that single messages can be classified with reasonable accuracy, but that the accuracy improves significantly when multiple messages in a chat are combined.

For classification of chats we have shown that the use of PCA for reduction of the dimension of the feature vector and the removal of dependency between features has a positive effect on the performance. We have considered various methods for combining the classification of the separate messages and the method that performed best was combining the messages based on the output score of the SVM model, where the performance reached 93% with a training size of 25. Higher accuracies were also obtained, but only for cases where the amount of test data was rather low.

Overall, we can conclude that we can, with high accuracy, distinguish between a

chat where the chatter is honest and uses his own identity and a chat where the chatter has assumed an alternative identity.

7.7.3 Future Work

A first step in extending this research is to extend the dataset. This can be in the amount of data per participant (i.e. longer chats), but mostly the number of chat participants should be increased. Additionally, inclusion of other, text-related, features will most likely also help improve the performance results.

In a chat it is important to know if the person on the other side is honest or not. In case the person on the other side it dishonest, then a chatter might not want to continue chatting with that person. This then means that there needs to be a criteria that, *during* the chat, can indicate to a chatter the "level of honesty" of the chat partner. Developing such a criteria is part of our future work.

In our experiments, people were, seemingly, not experienced in lying. A topic for investigation would be to determine how well we could distinguish an experienced liar, who has experience in impersonating another person. In such a case we could distinguish between the detection of a person who assumes the same fake identity every time, or someone who takes "random" fake identities. Such a person could for example be an actor who has to play different roles while acting.

Acknowledgments

We would like to thank the participants in the experiment for spending their time to help us collect data.

Chapter 8

Article 3: Exploring Keystroke Dynamics and Stylometry Features for Gender Prediction on Chat Data

Li, G., Borj, P. R., Bergeron, L., & Bours, P. (2019, May). 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) (pp. 1049-1054). IEEE. .

8.1 Abstract

Online anonymity is considered as one of the great gifts of the Internet, but it also brings dangers to society, such as cybercrime, online sexual abuse and bullying, and love scams. Many people are fond of chatting online to make new friends, but how can they be sure that the person sitting behind the other computer is really the person they claim to be? By studying stylometry and keystroke dynamics features from chat data, it is feasible to reveal the actual gender of an online user. In this paper, we examined stylometry and keystroke dynamics features from chat data, and proposed a Random Forest based gender prediction approach by analyzing these features. In order to evaluate the effectiveness of the proposed approach, a data acquisition was conducted to capture the keystroke dynamics and text information when participants were chatting remotely via Skype. All participants were invited to chat freely on any topic they prefered in order to get to know each other. Based on our experimental result, the proposed approach achieved 72% prediction

accuracy by analyzing on this free-text data captured only in 15 minutes.

8.2 Introduction

Nowadays, people are absolutely dependent on the Internet for their daily life to do business, to entertain, to communicate, etc. One of the great gifts of the Internet is the online anonymity, which protects the user's identity from being stolen or misused by other parties [190]. However, online anonymity is a double-edged sword, which also favors the bad guys to commit cyber-crime, cyber-bullying, online sexual abuse, love scams, etc. Especially, online sexual abuse has made children extremely vulnerable to child predators, as children are more exposed on the Internet than ever before. Several tragedies have happened where young children were contacted and threatened by a sexual predator who normally associated with a fake profile. A recent example of the use of a fake profile was between a young Norwegian boy and a predator on a social media application. In this particular case the boy was extorted and this led to him taking his own life [191]. If an ongoing chat conversation would be analyzed and reveal the true identity details (such as gender or age) of the chat-partner to young Internet user, such tragedies might be prevented in the future. Therefore, in this paper, we focus on predicting the chatter's gender by analyzing the online chat data.

While discussing online chat data, there are two types of information that can be captured and analyzed: textual information typed by the user and keystroke dynamics (KD) information which represents the user's typing rhythm. Analyzing the textual information to recognize authorship is often referred to as stylometry analysis. This has been primarily studied for the purpose of forensic authorship analysis. According to the classification in [192], stylometry features can be categorized as: character-based (e.g. percentage of upper case symbols), word-based, syntactic (usage of punctuation), structure-based (e.g. average number of messages per paragraph), function words (to express mood or attitude), n-grams, etc. According to [193] can these stylometry features can be used for gender classification. Conventional stylometry analysis depends on a relatively long text, such as, a full paragraph or chapter from a book, or even the whole book and are difficult to apply to short texts like a text message sent via a social media application, or the 140 characters posted on Twitter [65, 69].

KD is a behavioral biometric which identifies or recognizes users based on the way they type on their keyboards. It can be collected unobtrusively without disturbing the user [186]. In KD we measure when a key is pressed and when it is released and from that we can determine various features, in particular the duration of a key (how long it is held down) or the latency between two keys (time between releasing one key and pressing the next one). Besides identification and authentication has

KD been widely used for soft biometric goals such as emotional state detection [139], age detection [84] and gender detection [89].

In this paper, we focus on developing and analyzing the stylometry and KD features which can be extracted from single message sent out from an online chat conversation. The remaining of this paper is organized as follows: Section 8.3 overviews related work on the various methods of gender detection, and section 8.4 will describe the details of feature extraction method and the proposed gender prediction approach based on KD features; Section 8.5 reports the data acquisition and parameter setting; Section 8.6 gives the experimental results and discussion; finally, conclusion and future work are discussed in Section 8.7.

8.3 Related Work

Applying KD to predict the gender has gained attention from various researchers. Giot and Rosenberger [81] presented a method for gender recognition using KD with more than 91% of accuracy. They used the GREYC keystroke benchmark database and extracted different features from each sample. These features were related to the press and release of two consecutive keys, in particular press-press time, press-release time, release-release time, and release-press time. They also investigated the combination of all 4 features. Support Vector Machine (SVM) was used as the learning method to classify the gender, and the best result was achieved by using the combination of the 4 timing features. Fairhurst and Da Costa-Abreu [89] tried to classify the gender of users in a social network environment using the same GREY keystroke benchmark dataset, and they used K-Nearest Neighbour (KNN), Decision Tree (DT), and Naive Bayes (NB) as classifiers. Furthermore they also explored combining these three classifiers using three fusion techniques: (1) Dynamic Classifier Selection based on Local Accuracy class (DCS-LA); (2) Majority Voting; and (3) Sum. It was shown that the DCS-LA method performed best for gender classification.

Pentel [82] tried to predict the gender and the age of the user from KD data features. He did not only use latency and duration information, but for example also information about making corrections. In addition to keystroke dynamics data did he also include mouse dynamics features. He applied Logistic Regression, SVM, KNN, and Random Forest (RF) for binary classification, and finally, he showed that RF performed best for age and gender prediction. For gender prediction he obtained an f-score of 0.73.

Recently, Tsimperidis et al. [83] achieved a good accuracy (95%) for gender classification by KD. They calculated the average values of keystroke durations and press-press (PP) latencies. Several classifiers such as RF, SVM, NB, Multi-layer

perceptron, and Radial Basis Function Network (RBFN) were tested. It was illustrated that RBFN had the best accuracy and had a good time complexity, while SVM showed the second best accuracy but it was the second slowest method for gender detection.

Tsimperidis et al. [84] created a database with 'IRecU', which is a free text keylogger, to predict the age of the users. He used Artificial Neural Network (ANN) to predict the age of a user, and it was shown that as he decreased the number of age classes to the three the performance became better than when useing four age classes. Pentel [82] tried to classify the 10-15 years old users against all others by using Keystroke Dynamics and mouse patterns; he used f-score as a performance metric, and among the different types of classification methods, RF achieved the best result for age detection with f-score 0.73, same as for gender recognition.

8.4 Proposed Approach and Its Features Investigation

In this section, we first describe the stylometry and KD features that could be used for predicting the user's gender. Secondly, a RF based classification gender prediction will be proposed by analyzing the KD features extracted from 20 letter bigrams.

8.4.1 Stylometry Feature Extraction

Online chatting is different from writing an essay or composing an email. Such tasks are normally carried out individually and provide relatively long time for thinking and preparation. Online chatting is an interaction between two (or more) users, each sitting behind their own computer and expecting to receive a response from the chat partner shortly after sending their own message. By considering natural characteristics of male and female, we think the response time from male users may be different (in particular shorter) than that of female users. Such shorter response time may also lead to more typing errors in terms of key deletions. Due to the limited amount of text involved in a short online conversation, conventional stylometry features (such as the frequency of preferred words by male or female) would not be effective in our analysis. Therefore, the following four stylometry features will be extracted from a short online conversation:

- **Average thinking time:** this is the time between two messages, in particular we measure the time duration between releasing the last key for one message and the time when the user presses the first key for the next message. Then we calculate the average value for all time durations generated from that user in an online conversation.
- **Ratio value of key deletion:** ratio value is calculated by the number of all

deleted keys divided by the total number of keys typed by that user in an online conversation. This ration can be calculated based on the knowledge of all keys typed by a user, including the number of times the backspace key and the delete key are used to remove keys from the final output.

- **Average number of letters in a word:** we count the number of letters in all words typed by the user in an online conversation, then calculate the average value to represent this feature. Note that the words with less than 4 letters are excluded.
- **Average number of words in a message:** the total number of words divided by the number of messages is the feature value.

8.4.2 Keystroke Dynamics Feature Extraction

KD features are extracted from the 20 most common letter bigrams in a English corpus [2]. These bigrams and their frequencies are listed in Table 8.1. From each bigram, we extract 6 KD features which have been proofed for their discriminability on user authentication by other researchers [194, 78, 195]:

- f_1 : the duration of first letter, which is the time elapse between pressing the first key and releasing that key.
- f_2 : the duration of second letter, which is the time elapse between pressing the second key and releasing that key.
- f_3 : the latency between first key released and the second key pressed, which is the time difference between releasing the first key and pressing the second key.
- f_4 : the latency between first key pressed and the second key pressed.
- f_5 : the latency between first key released and the second key released.
- f_6 : the duration of the first key pressed and the second key released.

According to these definition, features f_1, f_2, f_4 and f_6 are always positive, yet features f_3 and f_5 could be negative.

8.4.3 Keystroke Dynamics based Gender Prediction

The proposed KD based gender prediction approach consists of two stages: first stage is to predict the gender for each message; second stage is to merge the prediction results from all messages to produce a final prediction. Note that when we

Table 8.1: Selected bigrams for keystroke dynamics feature extraction and their frequency according to [2]

Bigram	Freq.	Bigram	Freq.	Bigram	Freq.	Bigram	Freq.
th	1.52	he	1.28	in	0.94	er	0.94
an	0.82	re	0.68	nd	0.63	at	0.59
on	0.57	nt	0.56	ha	0.56	es	0.56
st	0.55	en	0.55	ed	0.53	to	0.52
it	0.50	ou	0.50	ea	0.47	hi	0.46

classify a whole conversation, we mean in this paper the classification of each of the chatters, based on only the messages of that chatter in that conversation. So the actual conversation between two chatters is processed twice, once to determine the gender of one chatter, and next to determine the gender of the other chatter. For simplicity's sake we will just refer to classification of a conversation while we actual mean the classification of one of the chatters in that conversation

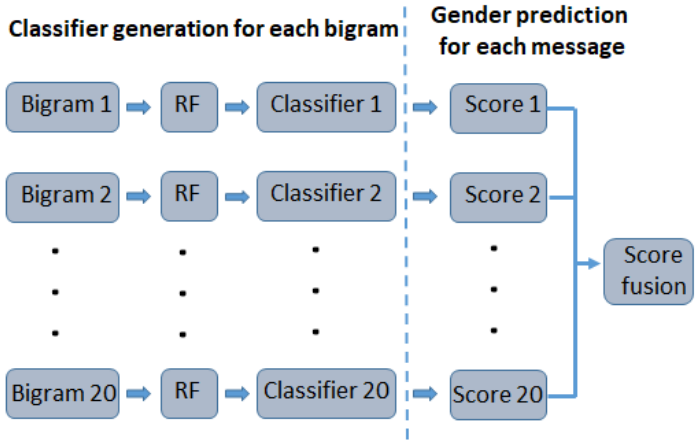

Figure 8.1: The diagram of a Random Forest (RF) based gender prediction approach by analyzing the keystroke dynamics features.

Figure 8.1 illustrates the core procedures of first stage prediction, which adopts RF as the classification technique to generate a classifier for each bigram. Each of the RF classifiers is trained to recognize 2 classes, 0 for female and 1 for male. The resulting classification score is value in the range $[0, 1]$, where a score less than 0.5 indicates female and a score over 0.5 indicates male. Each message contains a subset of the 20 bigrams that we consider in our analysis and each such bigram

will result in a score for the classification into male or female. To classify that message, score fusion is used. In particular is the median value of the bigram scores calculated. Here again, the message is classified as female if the resulting fusion score is below 0.5 and as male if the resulting fusion score is above 0.5. We did also test with using the mean of the bigram scores for the score fusion, but we found that using median gave a better performance.

After obtaining the fusion scores of each of the messages in a conversation, in the second stage the fusions scores and classification of each of the messages in that conversation are considered. Each of these messages has a fusion score as well as a classification as male or female. In order to classify the whole conversation, we perform simple majority voting (MV) on the classifications. For example if out of 16 messages 12 are classified as male and 4 as female, then the whole conversation is classified as male.

However, it is possible that, with an even number of messages in a conversation, half of the messages is classified as male and the other half as female. This tie can not be solved using simple MV. In this case we use the fusion scores of each of the messages. Assume that in a conversation n messages are classified as male, with fusion scores $(M_1, M_2, ..., M_n)$, where each M_i value is between 0.5 and 1. There are also n messages in that conversations classified as female and the fusion scores of these messages are denoted by $(F_1, F_2, ..., F_n)$, and now each F_i is between 0 and 1. We now calculate 2 values as follows:

$$Average_M = \frac{1}{n}\sum_{i=1}^{n}(M_i - 0.5) \tag{8.1}$$

$$Average_F = \frac{1}{n}\sum_{i=1}^{n}(0.5 - F_i) \tag{8.2}$$

Each of these averages will be an indication of how certain the first stage classifier was in determining that the message classified as male/female are indeed from a male/female chatter. A lower value indicates less confidence and a higher value more confidence. Keeping that in mind, we will (in case of a tie for MV) classify the full conversation as female if $Average_F > Average_M$ and as male otherwise.

8.5 Data Acquisition and Parameter Setting

8.5.1 Data Acquisition

In order to be as realistic as possible, we set up two identical computers at different locations and connected via Internet and we invited the participants who did not know each other from before. Each participant was asked to chat with other participant and get to know him or her. Each chat lasted about 15 minutes and Skype was used as it is a commonly used social media application. A logging tool called BeLT [80], developed in our laboratory, was used to capture all KD information during the chat. There was no topic specified for the conversation. Participants were asked to chat as normal as possible and try to get to know each other a bit.

8.5.2 Parameter Setting

There were 45 participants joining the data acquisition of which there are 10 female participants and 35 male participants. Even though this is an unbalanced dataset, we made sure that the amount of data for training the classifier was evenly balanced, such that there is no gender bias in the classifier output. This however means that the testset is highly unbalanced. In addition, the thresholds used in the evaluation will be cautiously selected in order to be neutral during prediction. For instance, the predicted gender will be male if the generated score S from the classifier is $0.5 < S \leq 1$, and the predicted gender will be female if such score satisfies $0 \leq S < 0.5$.

The proposed approach was implemented and evaluated in Python. We call the RF function from *sklearn* with 3-fold cross validation. During cross validation, 2/3rd of the data is used for training and the rest of the data is used for testing. In order to find the optimal values for the number of trees in the forest and the depth of tree, we developed a random search function with two parameters: the number of trees ranges from 200 to 2000, and the depth of three ranges from 100 to 500.

8.6 Performance Analysis and Discussion

8.6.1 Performance Analysis on Keystroke Dynamics Features

In order to confirm the discriminability of the selected 6 KD features, we list the average values for two bigrams (from Table 8.1) from 10 randomly selected male subjects and 10 female subjects. In Table 8.2, we can see the difference between male and female subjects.

Because of the uneven number of male and female participants, we were very cautious about the parameter selection to avoid biased training. During the classifier

Table 8.2: Selected bigrams for keystroke dynamics feature extraction and their frequency according to [2]

Bigram	f_1	f_2	f_3	f_4	f_5	f_6
th of male	79	70	38	117	108	187
th of female	70	75	67	138	143	213
ha of male	72	83	46	118	130	202
ha of female	78	95	60	139	156	234

Table 8.3: Gender prediction accuracy based on keystroke dynamics features with different amount of training data.

Num. of training subjects	Num. of testing subjects	Prediction accuracy
5 female and 5 male subjects	35	53.33%
10 female and 10 male subjects	25	72%

creation phase, the amount of KD features selected from the male subjects and female subjects are absolutely equal. Testing is done only with the data of participants that is not used during the training of the classifiers, i.e. the training and test dataset are disjunct. Table 8.3 gives the gender prediction accuracy by using different amounts of features during the training phase. As we can observe, the prediction accuracy can be significantly improved with a small increase of the training data. As this study was conducted on free-text KD analysis based on only 15 minutes conversation, we think that 72% accuracy is an encouraging result. We believe the accuracy could be boosted when we gain more data to build the classifier.

8.6.2 Data Analysis on Stylometry Features

The chat data was separated into message based data for generating the statistics of four stylometry features: average thinking time, ratio value of key deletion, average number of letters in a word and average number of words in a message. Note

that these four features were extracted from message based data, which indicates that the time for awaiting the response from other user is excluded for the statistics.

Tables 8.4 – 8.7 give the statistics in terms of median value, mean value and standard deviation regarding these four features. Figures 8.2 – 8.5 plot all feature values, where each mark represents a value calculated from one subject.

As we can see in Table 8.4, the median value for male's average thinking time is 393.7 ms while it is 506.5 ms for female subject. With respect to the ratio value of key deletion, the mean and median values for male subject are 0.0406 and 0.0382 respectively while they are 0.027 and 0.0268 for female subject. These considerable differences confirm our early assumption that these two features have the capability to distinguish the male from the female subjects in an online chat. This discriminability can also be observed from Figure 8.2 and Figure 8.3. However, we could not find the discriminability for the average number of letters in a word and the average number of words in a message as seen in Tables 8.6 – 8.7 and Figures 8.3 – 8.5.

Table 8.4: Median, mean and standard deviation of average thinking time (ms).

Gender	Median	Mean	StDev
Male	393.7	565.9	421.7
Female	506.5	591.1	615.3

Table 8.5: Median, mean and standard deviation of ratio value of key deletion.

Gender	Median	Median	StDev
Male	0.0382	0.0406	0.0189
Female	0.0268	0.027	0.0158

Table 8.6: Median, mean and standard deviation of average number of letters in a word.

Gender	Median	Mean	StDev
Male	4.0034	4.0079	0.22
Female	3.9898	3.9547	0.1579

timeAfterPunctuation.png

Besides the above statistics, these stylometry features were also used to form a feature vector for training the RF classifier. As the KD feature based classifica-

Table 8.7: Median, mean and standard deviation of average number of words in a message.

Gender	Median	Mean	StDev
Male	9.6667	10.0761	2.6963
Female	8.3523	8.9136	1.6134

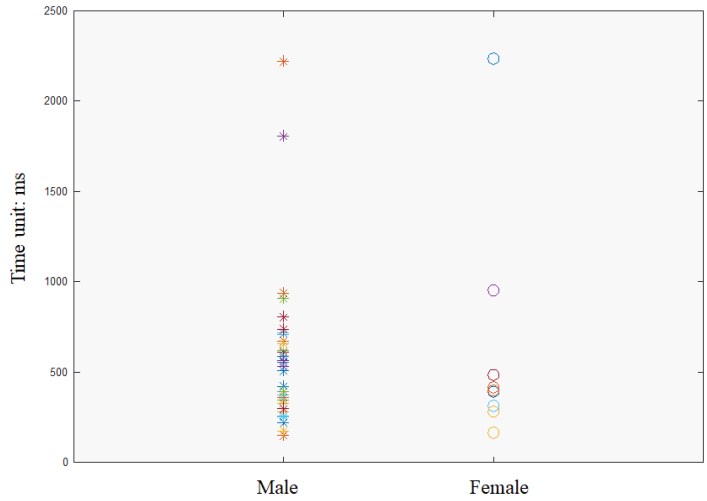

Figure 8.2: Average thinking time between two messages: each remark represents such average thinking time for one subject.

tion achieved the best prediction accuracy by using 10 male/female training subjects, we followed the same strategy and the prediction accuracy for stylometry features is 64% when the same amount of stylometry features extracted from 10 male/female training subjects were used to build the classifier.

8.6.3 Score-level Fusion and Impact of Messages' Length

After separately analyzing the KD features and stylometry features, it is interesting to consider the combination from these features. As the KD based classification is built on individual bigram features and not all bigrams can present in a short message, we analyzed a score-level fusion approach instead of feature-level fusion. The prediction accuracy from the score-level fusion approach remains at 64% with equal weights assigned to scores generated from the KD based classifier and the stylometry based classifier.

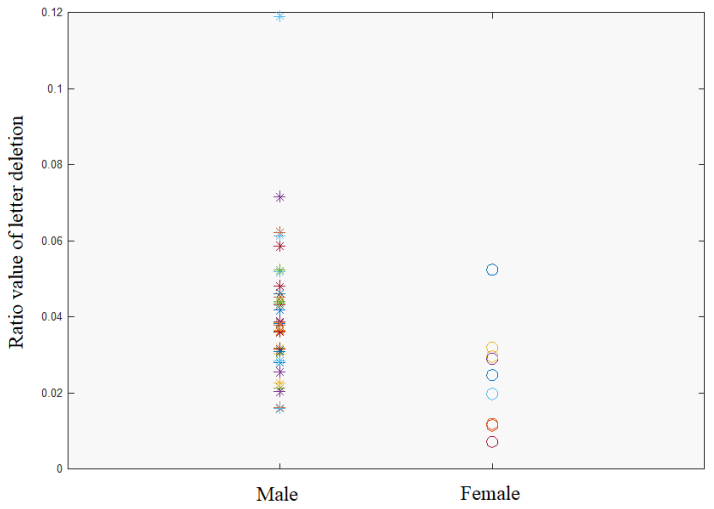

Figure 8.3: Ratio value of key deletion: each remark represents the ratio value of letter deletion for one subject.

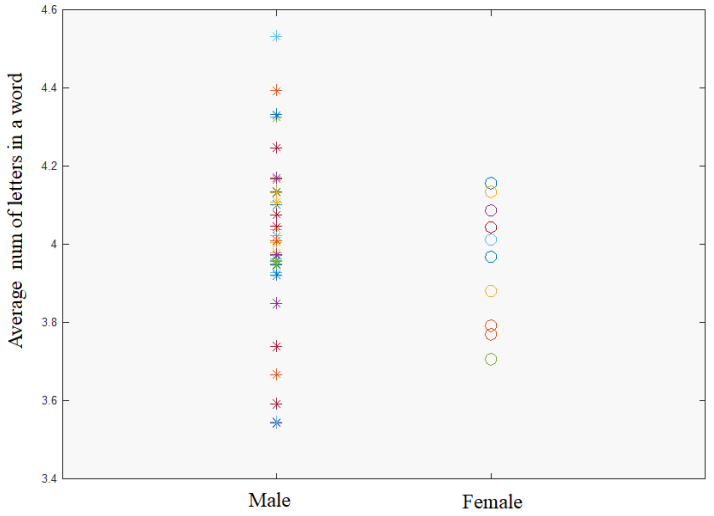

Figure 8.4: Average number of letters in a word: each remark represents such value for one subject, and this statistics excludes the words which has less than 4 letters.

As the length of message always varies and shorter message normally leads to less bigrams, it is also interesting to investigate the impact of messages' length from the performance perspective. Thus, we group the messages into three classes ac-

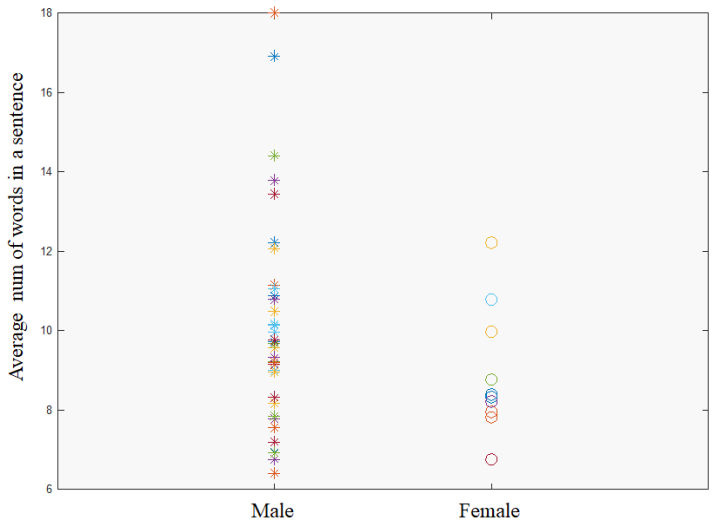

Figure 8.5: Average words of words in a message: each remark represents such value for one subject.

cording to the number of typed keys in the message: (!) short messages (less than 20); (2) intermediate messages (between 20 keys and 40 keys); and (3) long messages (more than 40 keys). Tables 8.8 and 8.9 give the prediction accuracy for KD features and for stylometry features respectively, when we consider the messages consisting of a certain number of keys. As we can observe, longer messages lead to better performance for the KD based approach, while the predication accuracy can be significantly improved when we only consider the message with more than 40 keys.

Table 8.8: Gender prediction accuracy based on keystroke dynamics features with different length of messages. These results are based on the training set consisting of 10 female and 10 male subjects

Length of message	Prediction accuracy
short message with less than 20 keys	52%
intermediate message with 20-40 keys	68%
long message with more than 40 keys	76%

Table 8.9: Gender prediction accuracy based on stylometry features with different length of messages. These results are based on the training set consisting of 10 female and 10 male subjects

Length of message	Prediction accuracy
short message with less than 20 keys	48%
intermediate message with 20-40 keys	44%
long message with more than 40 keys	84%

8.7 Conclusions and Future Work

In this paper we investigated the Keystroke Dynamics and stylometry features for predicting the gender by analyzing the chat data. A Random Forest based classification approach is also proposed to predict the gender by studying the KD features generated from 20 bigrams. In addition, we invited 45 participants to chat online with someone they never met before, for approximately 15 minutes, in order to evaluate the performance. The proposed RF classification was combined with a score-level fusion and a majority voting mechanism to predict the user's gender, and achieved 72% prediction accuracy. By analyzing the stylometry features on the collected data, we conclude that the pausing time between two messages and the letter deletion (or typo correction) have the capability to distinguish between a male and a female subject.

We think the presented work in this paper is a preliminary result, but revealed the feasibility of predicting the user's gender in a short online chat conversation. Regarding future work, there are many possibilities to explore and further improve the performance. Foremost we intend to extend the dataset with more participants, more data per participant (i.e. longer chats), while ensuring a more balanced gender distribution. Besides that we can also:

- consider the features from trigrams (instead of using only bigrams);
- consider more conventional stylometry features, such as the frequency of selected words, or words preferred by male or female, the number of punctuation, etc.;
- design different fusion mechanism when fusing the scores at message-level, as length of message affects the performance.

Predicting gender can be used to warn a chatter that the gender in the profile of the chat partner is not correct, for example a male subject trying to impersonate a female. In such a case, the chat partner of this impersonator would like to be warned as soon as possible about the incorrectness of the gender and not only find this out when a chat conversation has ended. For this we would like to extend this work such that a system continuously predicts the gender, based on the current and all previous messages received. Clearly, such a continuous gender prediction system needs to reach a certain level of assurance before a potential deviation between the claimed and determined gender is reported.

Chapter 9

Article 4: Predatory Conversation Detection

Borj, Parisa Rezaee, and Patrick Bours. "Predatory conversation detection." 2019 International Conference on Cyber Security for Emerging Technologies (CSET). IEEE, 2019.

9.1 Abstract

Providing a safe environment for children in online networks can be challenged by the anonymous nature of the internet. One of the worst forms of cyber-security issues is child grooming, where sexual predators seek contact with minors to abuse them. Various types of organizations such as chat providers and law enforcement are inclined to find online predatory conversations in order to protect the children. This paper proposes a study on predatory conversation detection using Natural Language Processing. We analyzed the different types of features of online grooming data considering various characteristics of online conversation, such as psycho-linguistic patterns. Our experiments with online communication showed an accuracy of 0.98 in automatically classifying the conversations into predatory conversations and non-predatory conversations. Best results are obtained by using linear SVM on 1-gram features when removing stop words as well as by using multinomial Naïve Bayes on 1-gram features when not removing stop words.

9.2 Introduction

In Greek language, pedophilia is an expression for love (philia) of young children (pedeiktos) [196]. Pedophilia is an expression that shows fantasies, sexual arousal, and sexual interests in children. The important characteristic of this definition of

pedophilia is age, as pedophiles intend to have a sexual relationship with minors [196]. In this paper we will use the term sexual predator (or just predator) to indicate a pedophile according to the above characteristics. It should be noted that a predator approaches the victim to build not only sexual but also emotional relationship, and the tactics used by child offenders for abusing the children are not the same [15]. So, finding sexual predators is a complicated task, and consequently, the number of children are affected by this predatory behavior is increasing as children have access to the internet through cellphones easily, and the internet provides a good opportunity for predators to cover their real identity using the anonymity characteristics of the internet.

Sexual abuse of children is an important and common issue in society, while many people, and especially parents, are not aware of this phenomenon. This might reflect that there is no specific definition and condition in different societies about child abuse or grooming. Craven [15] defined the grooming as *"a process by which a person prepares a child, significant adults and the environment for the abuse of this child."* Grooming has specific stages including gaining access to a child, gaining compliance, and maintaining the child's secrecy to avoid disclosure. This process serves to strengthen the offender's abusing pattern, as it may be used as a means of justifying or denying their actions. It affects the victim's life psychologically, physically, emotionally, behaviorally, and psycho-socially [15]. Grooming detection is a multifaceted and complex problem due to its variation in duration, type, and intensity depending on the perpetrator characteristics and behavior; identifying where and when the grooming process begins and ends might be impossible.

The need for sexual predator detection has increased by the development of the internet and online social networks. This development made the world more dangerous for children, as parents can hardly monitor their online activities. Consequently, protecting the children in an online world is a big issue. Social media has aroused harms such as addiction, depression, political polarization, and various types of cyber-criminal acts like sexual abuse of children. It is common to engage in various kinds of social media activities while it is easy to access the internet through smartphones. This makes it easy to have social interactions between children and adults, and therefore, it brings many risks for children in online communications. Predators try to gain the confidence of their victims in social media and then abuse them. For this reason, the detection of online sexual predators has gained interest of many researchers.

One of the main goals of Natural Language Processing (NLP) systems is to understand the meaning of what is being communicated. A lot of work has been done in traditional language documents such as books and longer articles. Online conver-

sations are more challenging to be processed by NLP as in most cases, they do not follow the standard grammar of writing and have a lot of slang words that might not be defined in the traditional languages. Besides, online chats are short, and they do not provide as much information as static texts such as books. Little work has been done in processing online conversation and specifically sexual predator detection in chat rooms. Egan [93] showed that finding the certain emotions in text could be helpful for predator detection in social media.

During the last decade, more attention has been given to face the cyber-pedophilia problem with different goals, including

- Identification of predatory chat lines;
- Classification of predatory chat conversations;
- Identification of the offender and the victim in a predatory conversation.

The Sexual Predator Identification competition (PAN2012) [30] was held in order to cope with these goals. Researchers proposed different methods to detect cyber-criminal activities. The competition included two parts, identifying predators, and predatory lines detection. The first part has two steps including predatory conversation detection and detecting the predators in predatory conversations. In this work, we focus on predatory conversation detection.

The remainder of this paper is organized as follows. In Section 9.3, we will first give a short overview of the State of the Art of the features that were used for predatory conversation detection. In Section 9.4, we will explain the data that is used and layout how the preprocessing is done, which type of features are extracted, and what the metrics are for performance evaluation, while in Section 9.5, we present the results of our analysis. Finally conclusions and future research directions are discussed in Section 9.6.

9.3 State of the Art

One of the first attempts for identifying online sexual predators was performed by Pendar [16] using K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM). He recognized the sexual predators and the underage victim with F-score of 0.943. In 2012, the PAN competition was held in order to find the best method for detecting the sexual predators among the chat messages, and also, finding the lines in predatory conversations which are the most distinctive for online grooming.

Using SVM, Villatoro et al. [61] obtained the best accuracy for predators identification. They tried to identify sexual predators in a set of suspicious chats conversa-

tions. They assumed that words used in online child abuse cases are different from general conversations, and also believed that predators use the same approach to catch their victims. To detect the predators, they performed two tasks: (1) Suspicious Conversations Identification (SCI) and (2) Victim From Predator disclosure (VFP). The accuracy of 0.988 for SCI and 0.925 for VFP were achieved by using the bag of words (BOW) representation employing either a Boolean or a Term Frequency - Inverse Document Frequency (TF-IDF) weighting scheme.

Parapar et al. [143] used Linguistic Inquiry and Work Count (LWIC) features in addition to TF-IDF features, and they got an F-score of 0.849 for the identification of sexual predators. Cano et al. [40] used a dataset which was based on chat conversation transcripts extracted from the Perverted Justice (PJ) website to classify chat messages into various grooming stages utilizing a series of features including sentiment polarity, content, and psycho-linguistic and discourse patterns. They acquired an F-score of over 0.8 to group the steps in the conversations.

Ebrahimi et al. [52] detected predatory chats using an anomaly detection method. They achieved an F-score of 0.915 to classify the chat conversation into predatory and non-predatory conversations. Recently, in [53] a Convolutional Neural Network (CNN) was used to mitigate the problem of predatory detection in chat rooms. It was shown that general pre-trained word vectors are not suitable for this problem, and not removing stop words, numbers and symbols during the analysis can increase the classification performance. Several machine learning techniques such as SVM, Neural Networks (NN) and CNN were used, the best F-score of 0.809 was achieved when using CNN [53].

9.4 Method

9.4.1 Data

Most work in NLP analysis has been done in formal documents such as essays of people, the Reuters news corpus, and the British National Corpus (BNC). Informal writing like blogs, tweets, and chats, have different characteristics: they contain spurious information, and more likely to contain grammar and spelling errors, abbreviations, slang words and phrases, and emoticons. Clearly, analyzing informal writings is more challenging than formal ones. Another crucial challenge in cyber-pedophilia detection is collecting relevant data. There are two different types of online text chat with sexual content

- Conversation between a sexual predator and the minor
- Conversation between two consenting adults

The fact is that chat providers do not make online conversations publicly available, as it requires the informed agreement of the participants, so access to this type of data is difficult while significant privacy and legal issues need to be resolved.

Most of the research regarding online sexual predators has used the conversations that are made available on the Perverted Justice (PJ) website (http://www.perverted-justice.com/) as a source of data for cyber-pedophilia detection. The Perverted Justice Foundation, more commonly known as Perverted-Justice (also known as PJ or PeeJ), is an American organization based in California and Oregon. Policemen and adult volunteers working for PJ, pretend to be minors in online social media, to serve as a decoy to attract the sexual predators.

Figure 9.1: XML file contains the conversation-ids and message information: author, time, text.

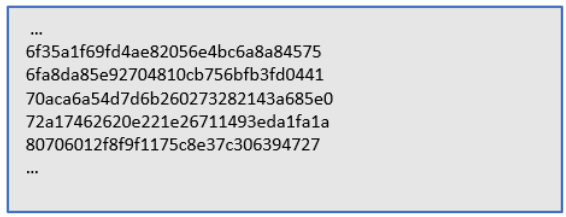

Figure 9.2: Predators IDs

The PAN2012 competition released a data set that can be used as a common reference point for research in sexual predator detection. The data is retrieved from various sources and contains: (1) normal (i.e. non-sexual) chat conversations; (2) sexual conversations between consenting adults (from OMEGLE); and (3) predatory conversations from PJ. Researchers can use their analysis techniques from different (e.g. Information Retrieval, Natural Language Processing, and Text Mining with Machine Learning) to train and test their systems and compare the performance of their method to other researches. Performance of a system is related to the number of True Positives (TP, i.e. correctly identified sexual predators or predatory conversations), True Negatives (TN, i.e. correctly identified non-predators or non-predatory conversations), False Positives (FP, i.e. non-predators classified as predator or non-predatory conversations marked as predatory), and False Negatives (FN, i.e. predators classified as non-predators, or predatory conversations marked as non-predatory). From these four values we can calculate the following performance indicators:

- **Accuracy:** The accuracy is defined as the fraction of correct decisions, i.e. $Acc = \frac{TP+TN}{TP+FP+TN+FN}$.
- **Recall:** Recall is the fraction of detected relevant items, i.e. $Rec = \frac{TP}{TP+FN}$.
- **Precision:** Precision measures the ration between the number of relevant detected items and the number of detected items, i.e. $Pre = \frac{TP}{TP+FP}$.
- **F_β-score:** The F_β-score is a weighted average of Precision and Recall, and in particular:
  $F_\beta = (1+\beta^2) \cdot \frac{Pre+Rec}{\beta^2 \cdot Pre+Rec}$.
  In case of equal weighting (i.e. $\beta = 1$), we can refer to either F_1-score or simply to F-score.

In this paper, we have used the data which was created for the PAN2012 competition [30]. Different data was made available for training and testing. The main

files contained the conversations between the chatters. The conversation file was given in XML format (see Figure 9.1 for an excerpt) contains a lot of conversations from various chatters. Each conversation has a unique conversation ID, and within the conversation are different chatters identified with different chatter IDs. A chatter can participate in multiple conversations. A conversation contains at least one message and hence one chatter, but most conversations have two or more chatters included. Each message in a conversation contains, beside the identity of the chatter, also the time of sending the message as well as the actual text message that was sent. A conversation where one of the chatters is a sexual predator is referred to as a "predatory conversation", while the remaining conversations are referred to as "normal conversation" or "non-predatory conversation". The PAN2012 data set also contained a text file listing all the chatter IDs of the predators (see Figure 9.2 for an excerpt), and this information can be used to label each of the conversations as predatory or non-predatory.

9.4.2 Preprocessing

As it was mentioned before, the conversations were tagged into two groups, non-predatory conversations, and predatory conversations. There are 64911 non-predatory conversations between users that might have more or less than two participants for chatting. There are 2016 predatory conversations between sexual predators and victims, and these conversations do have one or two chatters. All the conversations in the XML file that have only one chatter or have more than two chatters were removed. In other words, only the conversations between two persons were extracted for analysis. It should be noted that the data is highly biased, because the number of non-predatory conversations is much higher than the number of predatory conversations.

Some common words such as 'is' and 'the', which appear frequently in a text, do not provide enough valuable information in text mining while in some cases they might be valuable. Such words are referred to as stop words. In most text analysis work, these stop words have been removed, but it should be considered that online conversations such as chat and tweets are relatively short, in which case every single word can be substantial for text analysis, including stop words. To evaluate the usefulness of stop words in a chat conversation, we analyzed the data both with stop words and without stop words. We found that the performance when removing stop words is slightly better, except when using multinomial NB.

As it was mentioned before, online conversations do not follow the standard grammar. For instance, there are a lot of misspelling, slang words, and abbreviations in online texts. Some research works, in preprocessing step, have removed all the misspelling and have tried to turn the non-standard online conversations to the

standard text, but we feel that converting the online conversations to the standard language will loose a lot of relevant information that can give clues about the author of the text. So, to keep the information, we did not use any stemming or lemmatization during the preprocessing of the data.

9.4.3 Feature Extraction

A numeric feature space is needed to train the machine learning algorithms, which is an input with a two-dimensional array where rows are instances and columns are features. So, we need to transform the text data into vector representations which give the ability to perform meaningful analytic on text documents. It should be noted that a vector representation of a text can become extremely sparse, especially as vocabularies get more abundant, and it can have a significant impact on the performance of machine learning.

One of the common ways for extracting the features from a text is Bag Of Words (BOW) [52], and the decision is made based on the occurrences of a particular token in the text. It should be noted that the BOW method loses the relationship between the words as it is assumed that the text is represented as a bag (multi-set) of its words without considering the grammar and even word order. In this work, we not only count the token occurrences but also considered the token pairs, triplets, or different combinations. Here a token can be the representation of a word or a stop word. This approach is also called as extracting n-grams, and a 1-gram stands for separate tokens, 2-grams stand for token pairs, and so forth.

High-frequency tokens are stop words, and it was mentioned before we analyzed the data with and without considering them. In BOW, the goal is to find the tokens that are more discriminative between texts.

Several ways have been used for data representations including binary representation, Term Frequency or the number of occurrences, and TF-IDF [52]. In this work, the features were extracted based on the bag of words (BOW) models using the TF-IDF weighting scheme.

TF-IDF

The measure of term specificity first time was proposed in 1972 by Jones [197] that paper later became known as inverse document frequency or IDF [132]. The main idea is that *"a query term which occurs in many documents is not a good discriminator and should be given less weight than one which occurs in a few documents, and the measure was a heuristic implementation of this intuition"* [132]. The formula of TF-IDF is defined as below [198]:

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

where $w_{i,j}$ is the weight for term i in document j, N is the number of documents in the collection, $tf_{i,j}$ is the term frequency of term i in document j and df_i is the document frequency of term i in the collection [198].

A high weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of that term in the whole collection of documents.

Psycho-linguistic Schema

Various research works in authorship profiling have shown that when people write about a specific topic, they show different styles of writing based on their personalities and experience [40]. In addition, it was shown that sexual predators might suffer from emotional and instability and psychological problems. So, considering the sentiment polarity analysis might give us some clues about the online predators [199]. For example, to get the confidence of the victims, predators use more emotional words and emoticons. In some research work [40] [36], building a confidential relationship with the victim is called Deceptive Trust Development. Predators talk about their hobbies and what they like or dislike to befriend the minors.

As mentioned, predators seek to develop a confidential relationship with the minors, and in order to get their trust, they give a lot of compliments about the victims. Nuria [145] has defined the compliment as a *"speech act which explicitly or implicitly attributes credit to someone other than the speaker, usually the person addressed, for some 'good' (possession, characteristic, skill, etc.) which is positively valued by the speaker and the hearer."*

According to research, the compliment topics are mostly about appearance, ability/performance/skills, possessions, and personality. Non-linguistic variables such as the age and gender define the order of frequency of these compliments while apparently, appearance has been the main topic of complementing [145].

As predators are conservative in building the relationship with minors, they do not use hardcore words while adults might use this type of words in their conversations. Predators do not want to lose the trust of the children, and they assess the risk of their plans by making sure that the minor agrees with their ideas. For instance, they will ask many questions such as "would you like to ...?", "do you like ...?", "what do you like?", or "what do you want?". Moreover, it should be noted that sexual predators mostly try to define the topic of the conversations as they are

willing to gain enough information about the victim and assess the risk. They seek this goal by asking many questions. Below you can see some typical questions from predatory conversations.

is ur mom asleep?
i wanna see u do u think that is bad?
do your friends come over and hang out?

We compared the amount of questions between predators and victims, and we found that predators ask more questions than victims. In Figure 9.3 we can see that the average number of questions asked by a predator is higher than that of a victim. But we can also see that a chatter in a non-predatory conversation (tagged as "user 1") generally asks very few questions. This indicates that the number of questions in a conversation could be a weak indicator of it being a predatory or a non-predatory conversation. We have not used this however in our further analysis, because of the large spread in the number of questions asked (see the standard deviations displayed in Figure 9.3).

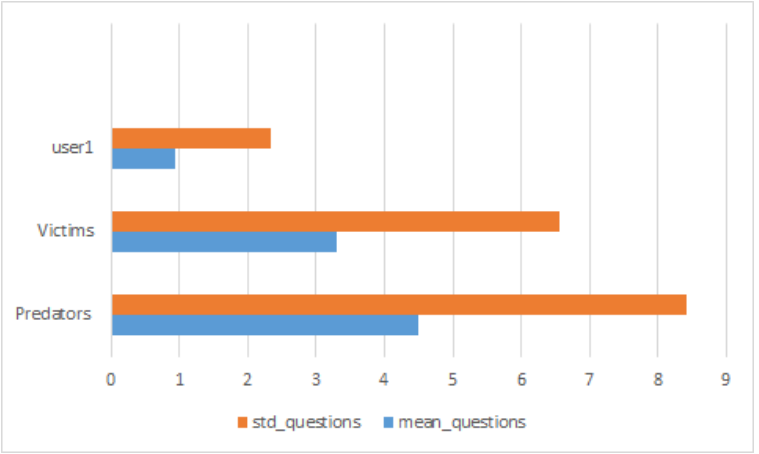

Figure 9.3: Comparison of the number of the questions which were asked

9.5 Results

We used the extracted features which are based on the BOW models, using TF-IDF weighting scheme to train the classifiers. The goal is to recognize the predatory conversations from normal, non-predatory chats between adults. The table shows the result of predatory conversation detection with various classification methods. The predatory conversation is assumed as a positive result. We trained the machine

learning methods considering three different n-grams including 1-grams, 2-grams, and 3-grams. Analysis was performed with SVM (both linear and non-linear), Random Forest (RF), and Naïve Bayes (NB). The results are shown in Tables 9.1-9.12. Each table shows the achieved accuracy on the whole testing part of the PAN2012 dataset, but also shows the obtained precision, recall and F-score on only the predatory conversations (Pos) and the non-predatory conversations (Neg). Note for example that although the accuracy in Table 9.5 is acceptable at 0.95, but the performance on the predatory conversations is terrible.

Accuracy = 0.98	Precision	Recall	F-score
Pos	0.91	0.82	0.86
Neg	0.99	1.00	0.99

Table 9.1: Linear-SVM model with 1-gram features

Accuracy = 0.98	Precision	Recall	F-score
Pos	0.87	0.84	0.86
Neg	0.99	0.99	0.99

Table 9.2: Linear-SVM model with 2-gram features

Accuracy = 0.98	Precision	Recall	F-score
Pos	0.81	0.85	0.83
Neg	0.99	0.99	0.99

Table 9.3: Linear-SVM model with 3-gram features

Accuracy = 0.96	Precision	Recall	F-score
Pos	1.00	0.25	0.40
Neg	0.97	1.00	0.98

Table 9.4: Non-Linear-SVM model with 1-gram features

Accuracy = 0.95	Precision	Recall	F-score
Pos	0.00	0.00	0.00
Neg	0.95	1.00	0.98

Table 9.5: Non-Linear-SVM model with 2-gram features

In text analysis with machine learning, one has to deal with very large number of features, often more than 10000 features. SVM uses protection against overfitting, which is independent of the number of features. Because of this, SVM has the

Accuracy = 0.95	Precision	Recall	F-score
Pos	0.00	0.00	0.00
Neg	0.95	1.00	0.98

Table 9.6: Non-Linear-SVM model with 3-gram features

Accuracy = 0.96	Precision	Recall	F-score
Pos	1.00	0.12	0.22
Neg	0.96	1.00	0.98

Table 9.7: Random Forest model with 1-gram features

Accuracy = 0.95	Precision	Recall	F-score
Pos	1.00	0.01	0.01
Neg	0.95	1.00	0.98

Table 9.8: Random Forest model with 2-gram features

Accuracy = 0.95	Precision	Recall	F-score
Pos	0.00	0.00	0.00
Neg	0.95	1.00	0.98

Table 9.9: Random Forest model with 3-gram features

Accuracy = 0.98	Precision	Recall	F-score
Pos	0.95	0.65	0.77
Neg	0.98	1.00	0.99

Table 9.10: Multinomial NB model with 1-gram features

Accuracy = 0.97	Precision	Recall	F-score
Pos	0.99	0.36	0.53
Neg	0.97	1.00	0.99

Table 9.11: Multinomial NB model with 2-gram features

Accuracy = 0.95	Precision	Recall	F-score
Pos	0.99	0.11	0.19
Neg	0.96	1.00	0.98

Table 9.12: Multinomial NB model with 3-gram features

ability to handle large feature spaces. From Tables 9.1-9.6 we can see that linear SVMs perform better than non-linear SVMs (denoted as SVC in Figures 9.4 and 9.5) when we have a high number of features [200]. From our results, we can also conclude that linear SVM outperforms RF and NB.

As it was mentioned before, while vocabularies get larger, the vector of features becomes sparser, which influences the performance of machine learning algorithms. Some machine learning methods such as Gaussian NB do not perform well on sparse data, and we found that multinomial NB is better suited for vectorized text data. We see from the various tables that 1-gram features result in a better accuracy than 2-gram or 3-gram feature sets in all machine learning methods on the used dataset. Using 1-gram features, linear SVM and multinomial NB have good accuracy while linear SVM gives a better F-score of 0.86 for predatory conversation detection than NB with an F-score of 0.77.

As aforesaid, to keep all the information, we performed the models with stop words and without them. The results given in Tables 9.1-9.12 are calculated when removing the stop words. Tables 9.13-9.16 show the result with stop words included. Here all analysis are performed on 1-grams as that gave the best results so far.

Accuracy = 0.98	Precision	Recall	F-score
Pos	0.89	0.80	0.84
Neg	0.99	1.00	0.99

Table 9.13: Linear-SVM model with 1-gram features keeping the stop words

Accuracy = 0.96	Precision	Recall	F-score
Pos	0.99	0.12	0.21
Neg	0.96	1.00	0.98

Table 9.14: Non Linear-SVM model with 1-gram features keeping the stop words

Accuracy = 0.95	Precision	Recall	F-score
Pos	1.00	0.02	0.03
Neg	0.96	1.00	0.98

Table 9.15: Random Forest model with 1-gram features keeping the stop words

Linear SVM and Naive Bayes work well keeping the stop words, and it should be mentioned that Naive Bayes performs better when using stop words compared to removing them. Linear SVM detects the predatory conversations with an F-score of 0.84, and NB recognizes the predatory conversations with an F-score of 0.83, and both SVM and NB have accuracy of 98% for predatory conversation detection.

Accuracy = 0.98	Precision	Recall	F-score
Pos	0.92	0.76	0.83
Neg	0.99	1.00	0.99

Table 9.16: Multinomial NB model with 1-gram features keeping the stop words

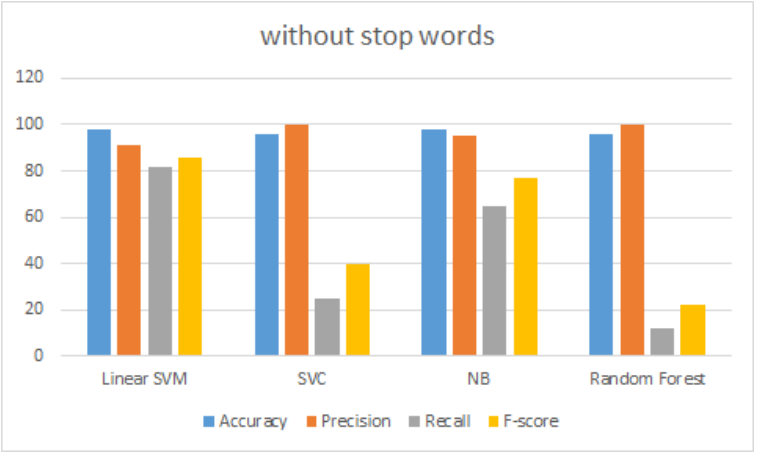

Figure 9.4: Comparison of predatory conversation detection methods

Although it was claimed that keeping the stop words is more beneficial than removing them, Figures 9.4 and 9.5 show that to detect the predatory conversations, it is better to remove the stop words as the classification methods gave better results, while they work without stop words. Linear SVM and multinomial NB outperform the other approaches for predatory conversation detection in both cases. Both of them classified the conversation with an accuracy of 98% while linear SVM has slightly better F-score of 0.86 to detect the predatory conversations.

9.6 Conclusions and Future Work

We performed different classification methods to recognize the predatory conversations from typical chats between adults. To build a confidential relationship, sexual predators try to identify their victims' needs and give a lot of compliments about different topics such appearance, ability, and personality. To this end will predators ask many questions to gain suitable information and seek their goals. It was shown that the amount of questions a predator asks is more than the number of questions a victim or non-predatory users might ask in an online conversation.

Our experiments show that preprocessing the data for predatory identification is quite beneficial. In text analysis, there are a lot of features, and SVM is a suitable

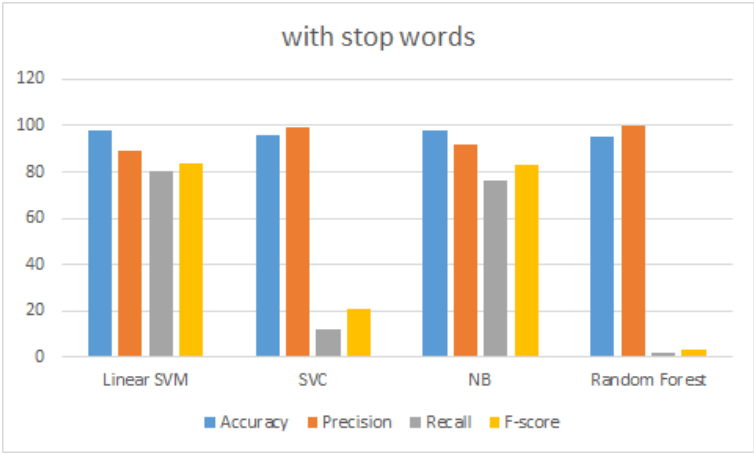

Figure 9.5: Comparison of predatory conversation detection methods

classification method to handle data sets with large sparse feature spaces, and more precisely, linear SVM is a good approach for data analysis with high dimensional feature sets.

Various types of classification techniques were used to detect predatory conversations, including linear SVM, SVC, Naïve Bayes, and Random Forest. Linear SVM and NB recognized the predatory conversations with an accuracy of 98% although linear SVM gives a better F-score of 0.84 for predatory talk detection.

Trying different n-gram feature sets showed that 1-gram is most suitable for BOW features to detect grooming conversations. Although BOW gives a good result for predatory conversation detection, it should be noted that BOW features do not consider the relationship between the words, and in future, we have a plan to include the features extract the relations of the words and sentences for this problem.

Future work will also include further analysis of predatory conversations where we will investigate the correct determination of the predator and victim in a predatory conversation.

Chapter 10

Article 5: On Preprocessing the Data for Improving Sexual Predator Detection

Borj, Parisa Rezaee, Kiran Raja, and Patrick Bours. "On preprocessing the data for improving sexual predator detection: Anonymous for review." 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA. IEEE, 2020.)

10.1 Abstract

Sexual predator detection and predatory message identification are critical to avoid under-aged children from being abused online. In this paper, we investigate different feature extraction approaches for predatory detection. While the previous results indicate good accuracy on predatory conversation detection, there is a missing investigation on the robustness of feature space. Further, we also show the impact of preprocessing on data to improve the performance of predator identification and predatory message classification. Various types of the bag of words features, including binary, term frequency, and TF-IDF representation are investigated on the publicly available PAN 2012 competition dataset for predator identification. Further, to cover the relationship between the words in the text analysis, the GloVe feature set is also investigated for word embedding features. With the set of preprocessing of data, we illustrate the improvement in detecting predatory conversation with an accuracy of 0.994 and F_1-score of 0.964.

10.2 Introduction

Innovation in computer science and technology has advanced ways of communication. Various media, such as social media and messaging applications, have opened opportunities for public and private communication. While the public communication forums facilitate seamless open discussions and debates, spreading hatred and propaganda based opinion manipulation can be foreseen. However, such challenges can be addressed by having moderators to avoid misuse. In the case of private messaging, the challenge is highly critical, where the messaging cannot be moderated. Multi-party or two-party communication can be easily misused for criminal activities, planning cyber-critical attacks on governmental organizations and, in the worst case, abusing the under-aged children. The abuse of children can be for obtaining sexual favors, blackmailing for sexual acts, or even radicalizing political opinions. Pedophilia is defined as the sexual interest of adults in minors, and obviously, age plays a vital role in this aspect [196]. Given the critical and open nature of the problem, we restrict our focus on detecting sexual abuse through messaging in this work, which is broadly classified under pedophilia.

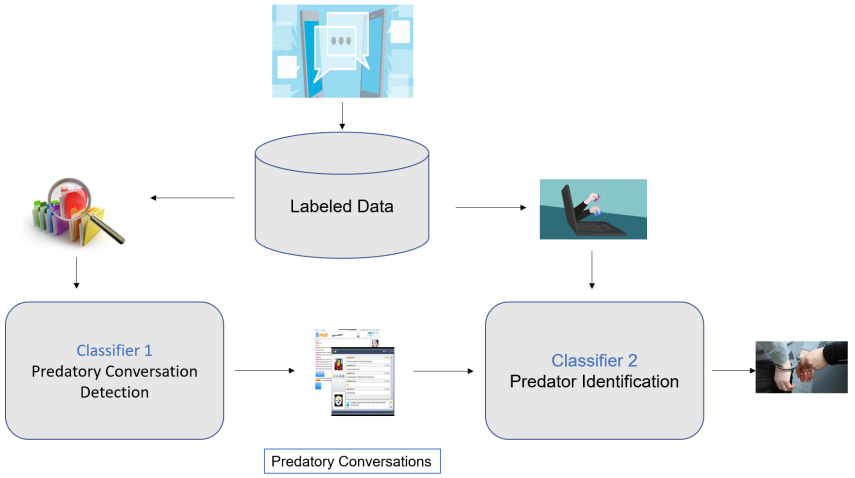

Figure 10.1: The Architecture of the Sexual Predator Detection System

Reformulating and adapting the concepts of pedophilia in the context of social communication, we refer to a *pedophile as a predator*, and any conversation that involves a pedophile is referred to as a *predatory conversation*. Therefore, it is critical to detect such behaviours in the messaging platforms at the earliest stage to avoid unpleasant experiences the under-aged children may undergo. However, detecting persons with pedophile intent is a challenging problem due to the complex

combination of underlying behaviour such as psycho-dynamics, differing motives, and style of writing depending on human characteristics. The considerable variation in these behavioural patterns makes detection of predators, particularly on online platforms, a complicated and hard problem [34].

Many predators might fulfill their needs through available online child pornography and will never try to get in physical contact with a child. However, with a carefully designed grooming[1] approaches and the advancement in communication technology, the chances of predators contacting the victims physically cannot be ruled out [34]. Whether-or-not, the child victim, is contacted physically, it is necessary to detect predatory conversations in the early phase to avoid the negative psychological impact on the under-aged victims. Depending on the predator's characteristics, grooming might have different stages, such as gaining access to the child, gaining compliance, and maintaining the child's secrecy to avoid disclosure. Predator and grooming detection is a complex problem, as every predator might have a specific way of approaching a child. While this can be to a certain degree handled by interception of chat messages by parents, in many cases, the parents may be completely unaware of predatory conversation leading to adverse impact on children, not only physically but also psychologically.

The magnitude of the problem of detecting the predatory conversations and eventually predators is increasing due to the limited text available during the chat conversations. Well established approaches of Natural Language Processing (NLP) can be used to analyze the textual information in online communications, such as detecting the topics. Despite the techniques in NLP, the varying style of writing and linguistics for communication hinder it from being effectively employed to detect predatory conversations. Short messages, usage of abbreviations and slang words, grammatical errors, and limited data available to identify the patterns make the problem highly complicated for efficiently detecting the predatory conversations.

While the meta-information such as age, gender, location can be easily used for identifying the predators, it has to be noted that the identity of users in online chat rooms is self-chosen, and many users will not provide their real age or gender. Precisely, with the intent of obtaining sexual favors, the predators adjust their age and gender to contact potential victims in the first place. While earlier work indicated the use of Keystroke Dynamics (KD) [25] to capture the specific typing rhythm of a user (i.e. not what text is typed, but how this text is typed) to identify the predators, most online social media or chat rooms have not integrated such an approach due to privacy issues related to the capturing of KD information, and the amount of data for transmission.

[1] A process that prepares the minor for sexual abuse in a particular environment is called child grooming [15].

Despite the limited data from the predatory conversations and non-availability of the meta-data of the predators, the problem of detecting predators is pressing. To generalize the problem, one can consider the problem of identifying the predators as a classical two-class problem where the predators are considered one class and the rest in another class. As can be deduced, this leads to a high class-imbalance problem, which is complex and needs robustness of various proposed approaches. In an attempt to provide feasible solutions for the same, much research has been focused in this direction [201], including the Sexual Predator Identification competition (PAN 2012) [30]. The challenge consisted of three main goals to be solved:

- Identification of predatory chat lines;
- Classification of predatory chat conversations;
- Identification of the predator and the victim in a predatory conversation.

In this work, we focus on similar goals in line with PAN 2012 with a primary purpose of investigating the textual features alone for predatory conversation identification and predator identification. The overall pipeline for predatory conversation identification and predator identification can be seen in Figure 10.1. In order to provide an objective answer on the role of different textual features, we provide an extensive analysis of four different features primarily used in NLP for the task of predator identification. We investigate Bag-of-Words with Binary representation, Term Count (TC) [202] and Term Frequency - Inverse Document Frequency(TF-IDF) [132], [198] while we also employ the deep learning-based text feature extraction by leveraging GloVe [134] word embedding. We bench-mark all the approaches with different established metrics on the PAN 2012 dataset and provide a detailed analysis. This work also provides insights on the most appropriate features, specifically to deal with the high class-imbalance and studies different classifiers to indicate the robustness of the models.

The remainder of this paper is organized as follows. In Section 10.3, we first give an overview of the State of the Art in online sexual predator detection, and we describe features that were used for predatory conversation detection. In Section 10.4, we provide the details of the dataset used, explain various preprocessing steps. In the same section, we also provide details on features extracted. In Section 10.5, we present the metrics used for performance evaluation, the results of our analysis and conclusions, and finally, future research directions are discussed in Section 10.6.

10.3 State of the Art

We list a set of relevant state-of-art approaches related to predator detection in this section. Egan et al. [93] tried to discover the language pattern used by sexual predators in chat rooms. They found various themes that sexual predators might express their emotions and patterns in an online conversation such as implicit or explicit content, minimizing the risk of detection, and preparing to meet offline. In their work, the authors noted that predators try to build an online relationship with the minors, and they mostly avoid any use of hardcore words to avoid losing any trust. At the same time, they assess the risk of their actions by asking many questions to gain enough information. To develop a confidential relationship, they explicitly or implicitly give compliments to the minors for appearance, ability/performance/skills, possessions, and personality [36], [145]. For instance, Cano et al.[40] used a combination of feature spaces including, bag of words (BoW), syntactical, sentiment polarity, content, psycholinguistic, and discourse patterns to detect the behaviour of child grooming in social media, and based on these features and their results, they have found several stages for online grooming that reveal the behaviour of predators.

Pendar [16] tried to identify online sexual predators using K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM). The features were extracted using word n-grams and document frequencies. These features were combined with Term Frequency - Inverse Document Frequency (TF-IDF) for his experimental result. Villatoro et al. [61] used a Boolean and a TF-IDF weighting scheme for sexual predator identification with an accuracy of 0.988 for predatory conversation detection. The authors detected the predators with an accuracy of 0.925. Parapar et al. [143] extracted linguistic inquiry and the bag of words (BoW) features, including word counting and TF-IDF. The approach identified sexual predators with an F-score of 0.849.

Ebrahimi et al. [52] used TF-IDF features with semi-supervised anomaly detection for the recognition of predatory conversations, which achieved the best accuracy of 0.99, with an F-score of 0.778. Bogdanova et al. [47] have used a set of high-level features for detecting cyber-grooming on the data that was extracted from the Perverted Justice (PJ) website (`www.perverted-justice.com/`). The features were based on the precise rate of emotional instability, such as feelings of inferiority, isolation, loneliness, low self-esteem, and emotional immaturity, and the best accuracy of 0.98 was achieved. Yun-Gyung Cheong er al. [64] proposed a method for predatory behaviour detection in online games. Various types of features were used, including BoW, Sentiment Features, and Rule-Breaking Features on the data that was extracted from MovieStarPlanet (`www.moviestarplanet.com/`). The

approach resulted in an accuracy of 0.93 and an F-score of 0.78.

10.4 Dataset and Feature Extraction

In this work, we employ PAN 2012 competition dataset [30] for all the evaluation. PAN 2012 competition dataset consists of the training set and testing set in a disjoint model. Both the training and the testing set contain a number of conversations. A unique conversation ID identifies each conversation and each message in a conversation contains an author ID, a timestamp of the conversation, as well as the text of the message. An author can appear in various conversations. Additionally, two files of the IDs of the sexual predators are given, one for training data and one for the testing data. In Figure 10.2 an example of the conversation format is presented, while in Figure 10.3, we can see part of the list of sexual predator IDs. The organizers of the PAN 2012 competition have collected the data from various sources. The sources included regular conversations without any sexual content, sexual conversations between consenting adults from Omegle (www.omegle.com), and conversations that are made available on the Perverted Justice (PJ) [2].

10.4.1 Pre-processing

As the sexual predators' IDs were given in the PAN 2012 competition dataset, in this work, we have labeled the conversations as predatory or non-predatory. A predatory conversation is a conversation where one of the parties has an ID that was in the list of sexual predator IDs. All the conversations where none of the parties' ID was listed of sexual predator IDs were labeled as a non-predatory conversation. Both training and test datasets contained conversations with one, two, or more than two authors. In this work, we focus only on conversations with two authors. Conversations with more than two authors are removed because sexual predators will contact potential victims in person, one-to-one conversations, and not engage in group chats. Conversations with only a single author are also not relevant as it does not represent a combination of a sexual predator and potential victim where cyber-grooming happens. Further, to curate the dataset, we also removed all conversations with less than seven messages as such conversations contain little information to be classified as either predatory or non-predatory.

Online conversations, in general, do not follow standard grammar, contain many misspellings and slang words, and be interspersed with emojis. The PAN 2012 competition data did confirm these findings to be right in their dataset[30]. The data contained a lot of non-English words that did not have any specific meaning,

[2]The Perverted Justice website is by an American establishment where police officers who work for PJ pretend to be minors in online social media to serve as a trap and attract sexual predators.

```
...
<conversation id="4823155b557200771ff561a3ba14969f">
  <message line="1">
    <author>ec37ddb0b8ce6dde7eb9114cd4077866</author>
    <time>12:58</time>
    <text>hi</text>
  </message>
  <message line="2">
    <author>988778820848f4c6123e604fad9a5c85</author>
    <time>12:58</time>
    <text>hhhhhhh</text>
  </message>
  <message line="3">
    <author>ec37ddb0b8ce6dde7eb9114cd4077866</author>
    <time>12:58</time>
    <text>asl</text>
  </message>
 </conversation>
 <conversation id="74890d7967c3bc8d4210d273938d155e">
  <message line="1">
    <author>0896b9b46ecc19b58af02d74761c049d</author>
    <time>18:17</time>
    <text>HELP. I have a html page that has a form
</message>
  <message line="2">
    <author>a293851ef5a495a148359d36fcb08260</author>
    <time>18:18</time>
    <text>you could use the onfocus event</text>
  </message>
...
```

Figure 10.2: Excerpt of conversations in PAN 2012 format

and that did not provide any information for training. All these kinds of non-meaningful words and symbols were removed from the text messages to study the real impact of feature space and classifier space. Further, we have converted the chat messages to standard texts where possible, and the text in the original form was retained in other cases. Converting the messages to the standard language is anticipated to result in the loss of some relevant information related to the author of the text. To overcome this and yet keep the information as much as possible, we do not carry out any stemming or lemmatization in the preprocessing of the data. The preprocessing procedure used for cleaning the data was the same for the training and the test dataset. The statistics of the dataset for predatory and non-predatory data in training and test sets are shown in Table 10.1.

	Predatory	Non-predatory
Training	951	8477
Test	1697	19922

Table 10.1: The number of conversations for each class

```
...
00851429b21722a4d62f63a328c601ca
00aac10b39157377c79b7700b7b832bf
02800e11fdb1b43595303709f2b38f8c
03957f443c7790f9642db14bbc59df11
...
```

Figure 10.3: Excerpt of predator IDs in PAN 2012 format

10.4.2 Feature Extraction

As recommended in the standard NLP approaches, the text data is converted to vector representations of numeric values. Many different approaches can be used to compute the vector representation of the data. In this work, we specifically focus on Bag of Words and Word Embeddings following the state-of-art works listed in Section 10.3. For the brevity of the reader's comprehension, we present a brief background of the feature extraction approaches.

Bag of Words

One of the traditional ways to determine a vector representation of a text is the Bag of Words (BoW) representation. The boW consists of a dictionary of words with all possible words. An index represents each word (also called a token) in the dictionary to a high-dimensional vector. A text is now represented in a sparse vector, where all entries are zero, except those words/tokens present in the text. The non-zero entries in the vector representation can be further coded in various ways as listed below:

1. **Binary:** In the Binary representation, a 1 value indicates that the word appears at least once in the text. Simple counting of the different occurrences indicates the weight of the vector.

2. **Term Count (TC):** Unlike simple binary counts, in Term Counts, the value accounts for the number of occurrences in the text, i.e., it represents a counter of occurrences. In this case, the weight of the vector equals the number of tokens in the text.

3. **Term Frequency (TF):** Term Frequency can be seen as a normalized version of the Term Count. The values represent a rational number between 0 and 1, representing the fraction of the words' occurrences. As the occurrences are fractions, the weight of the vector now equals 1.

4. **Term Frequency - Inverse Document Frequency (TF-IDF):** TF-IDF normalizes the term frequency (TF) values based on the occurrence in other

documents. If a term occurs in many documents, then it has lower discriminating power. This is compensated by using the inverse document frequency when weighing the terms to make it more discriminable.[132].

The computation of TF-IDF values is shown in Equation 10.1 [198].

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{10.1}$$

Here $tf_{i,j}$ represents the frequency of term i in document j, N represents the number of training documents, and df_i represents the number of documents containing term i. Finally $w_{i,j}$ is the value of term i when representing document j in the TD-IDF vector representation [198]. A high term frequency reaches a high weight in the TF-IDF representation (in the given document) and a low document frequency of that term in the whole collection of documents. In other words, the term is a-typical (i.e., not used in many documents), but typical for the author of the given document (i.e., a long term frequency).

It should be noted that BoW regards a text as a set of words, regardless of the grammar or even the word order. Clearly, it loses some information related to the relationship between the words. The primary goal in BoW is finding the most discriminative tokens. The BoW approach typically results in a larger dimension of feature vector depending on the size of the dictionary created. Generally, the dimensions are very high (typically 15-20.000), and further, the feature vectors are sparse, i.e., most values are equal to 0. BoW creates a so-called one-hot representation of a word/text. Following the state-of-art works, we employ binary, TC, and TF-IDF representations for the feature space analysis.

Word Embeddings

While BoW has been employed in earlier works, it has to be noted that in BoW, the relationships between words are entirely ignored, leading to loss of information. Such a loss of information leads to a sub-optimal classification of texts. For example, *"toy dog"* and *"dog toy"* have entirely different meanings but will be encoded in identical feature vectors when using BoW.

Word embedding is an alternative approach which mathematically transforms a high dimensional space into a (relatively) low dimensional space, in such a way that words with similar meaning will be encoded into similar feature vectors. **Word2Vec** [203] and **GloVe** [134] are examples of word embeddings recent used for extracting textual features. Despite the need to train on specific datasets for extracting efficient word embedding network, recent works have introduced generic pre-trained model[134]. Given the limited amount of data employed in this

work, we employ a pre-trained GloVe embedding model to obtain a feature vector of dimension 300 [134].

10.5 Results

We detail the evaluation and performance metrics employed in this work, followed by the details on results obtained on the dataset. We report two sets of experiments for predator identification and predatory conversation identification. Table 10.1 presents the details of the dataset used in this work. It has to be noted that the dataset is unbalanced, where the number of predatory conversations is much lower than the number of non-predatory conversations as detailed earlier.

10.5.1 Performance Measures

Given the two-class classification nature of the predator detection and predatory conversation identification, we employ the standard evaluation metrics of Accuracy (Acc), Precision (P), Recall (R), and F-score (weighted harmonic mean between precision and recall). Precision and Recall are calculated based on the number of True Positives (TP, i.e., correctly identified sexual predators or predatory conversations), True Negatives (TN, i.e., correctly identified non-predators or non-predatory conversations), False Positives (FP, i.e., non-predators classified as predator or non-predatory conversations marked as predatory), and False Negatives (FN, i.e., predators classified as non-predators or predatory conversations marked as non-predatory).

- **Accuracy:** The accuracy is defined as the fraction of correct decisions, i.e. $Acc = \frac{TP+TN}{TP+FP+TN+FN}$.
- **Recall:** Recall is the fraction of detected relevant items, i.e. $R = \frac{TP}{TP+FN}$.
- **Precision:** Precision measures the ration between the number of relevant detected items and the number of detected items, i.e. $P = \frac{TP}{TP+FP}$.

Precision is an indication of the number of false positives detected, while recall indicates the number of relevant instances not detected. While we note that standard F-score measure (called F_1-score) weights the precision and recall equally, in certain instances, it is relevant to focus either on the precision or recall. Specifically, in the case of the problem of detecting online sexual predators, it is relevant to focus on precision which is more relevant in order to not overload police with FP, i.e., investigating falsely accused non-predators. Thus the precision should be weighted lower, i.e., $\beta < 1$ in the definition of F-score below. Alternatively, the goal might be to reduce the number of false negatives to ensure that the number of non-detected sexual predators is as low as possible, i.e., recall being weighted

higher or $\beta > 1$ below. The measures of generalized F-score can be indicated as below: $F_\beta = (1 + \beta^2) \cdot \frac{P+R}{\beta^2 \cdot P+R}$.

In case of equal weighting (i.e. $\beta = 1$), we refer to F_1-score. When the focus is on precision, then we use $\beta = 0.5$ and refer to $F_{0.5}$-score, while for we use $\beta = 2$ and the F_2-score for more focus on recall.

10.5.2 Predatory Conversation Detection

We focus on this work on analyzing the feature space for the effective classification of Predatory Conversation. We, therefore, reproduce the state-of-art results first and then provide an additional analysis with a careful preprocessing from our work. To illustrate the preprocessing effectiveness, we compare the features extracted with BoW (TF-IDF, TC, and binary) and GloVe vectors. Earlier work also employed 1-grams to obtain the best result for predatory detection [26] and motivated by the earlier work, we employ 1-grams on BoW to extract the features. For all the set of features, we employ various classifiers such as SVM (both linear and non-linear), Random Forest (RF), and Naive Bayes (NB) to establish the relation of features.

For the detection of predatory conversations, we merged all the messages of a single conversation into a single text block and extracted the features from each of the merged texts. To create the training set, each merged text was labeled as a predatory conversation if one of the chatters in that conversation was identified as a predator, and as non-predatory otherwise. We have not applied feature selection on any of the BoW feature sets, as it was earlier shown that reduction of the feature space leads to a reduction in the detection performance[61]. Table 12.3 shows the results obtained for predatory conversation detection with various feature spaces.

We note a set of observation from the obtained results as following:

- As observed from Table 12.3, all types of feature spaces give a promising result using the SVM classifier (either linear or non-linear).
- As observed in PAN 2012 competition [30], it was noted that precision was more important than recall to not overload law enforcement authorities with many false negatives. In order to be consistent with the earlier results, we employ $F_{0.5}$-score as the criterion to evaluate the methods.
- Further, the F_2-score is also included to reflect higher weight on recall, i.e., the focus is on detecting as many sexual predatory conversations as possible.
- Considering the F_1-score or $F_{0.5}$-score, we observe the best results are obtained for the SVM classifier on the TF-IDF features. The highest F_2-scores

Feature	Classifier	Accuracy	Precision	Recall	F_1	$F_{0.5}$	F_2
GloVe	LinearSVM	0.989	0.95	0.91	0.930	0.942	0.918
	NonLinSVM	0.990	0.94	0.92	0.930	0.936	0.924
	RF	0.971	0.98	0.64	0.774	0.886	0.688
	NB	0.789	0.26	0.90	0.403	0.303	0.603
TF-IDF	LinearSVM	**0.994**	**0.99**	**0.94**	**0.964**	**0.980**	**0.950**
	NonLinSVM	**0.994**	**0.99**	**0.94**	**0.964**	**0.980**	**0.950**
	RF	0.937	1.00	0.20	0.333	0.556	0.238
Binary	LinearSVM	0.991	0.96	0.93	0.945	0.954	0.936
	NonLinSVM	0.991	0.96	0.93	0.945	0.954	0.936
	RF	0.931	1.00	0.12	0.214	0.405	0.146
	NB	0.993	0.99	0.92	0.954	0.975	0.933
TC	LinearSVM	0.990	0.95	0.92	0.935	0.944	0.926
	NonLinSVM	0.990	0.95	0.92	0.935	0.944	0.926
	RF	0.938	0.99	0.21	0.347	0.568	0.249
	NB	0.993	0.92	0.99	0.954	0.933	0.975

Table 10.2: Results for predatory conversation detection

are obtained with Naive Bayes on TC features.

Various experiments on benchmark datasets have shown that selecting the suitable combination of preprocessing methods concerning the problem's goal might provide a substantial difference in classification accuracy [204]. Motivated by this, we conduct a similar study in this work by investing the role of preprocessing on various features on the PAN 2012 dataset. As noted in Table 10.3, we can see that the preprocessing approaches achieve better results by refining the dataset to remove the unimportant data. It should be mentioned that the system in [61] reached first place in the PAN 2012 competition. We can see in Table 10.3 that even though the accuracy of the various methods is comparable, the F_1-scores show a significant difference. The $F_{0.5}$-score of Villatoro et al. [61] was equal to 0.935, the highest of all submissions, but much lower than the highest $F_{0.5}$-score of 0.980 in our results (SVM on TF-IDF features) simply achieved by better preprocessing approaches.

In a similar manner, we also investigate the word embeddings through proposed GloVe feature vectors using the different classifiers. We compare the obtained results of the GloVe features using SVM with the results from Ebrahimi et al.[53] using GloVe-CNN. As noted in an earlier set of experiments, the preprocessing method used for refining the dataset has generated better results as against earlier obtained results, as listed in Table 10.4.

Source	Features	Classifier	Accuracy	F_1
Villatoro et al. [61]	TF-IDF	SVM	0.988	0.952
Ebrahimi et al. [52]	TF-IDF	SVM	0.985	0.610
	TF-IDF	One-class SVM	0.982	0.546
Proposed method	TF-IDF	SVM	**0.994**	**0.964**
	Binary	NB	0.993	0.954
	Binary	SVM	0.991	0.945
	TC	SVM	0.990	0.935
	Glove	SVM	0.990	0.930

Table 10.3: Comparing the result with state of art results

Source	Feature	Classifier	Precision	Recall	F_1
Ebrahimi et al. [53]	GloVe	CNN	0.91	0.72	0.805
Proposed model	GloVe	SVM	**0.95**	**0.91**	**0.930**

Table 10.4: Comparing the result using GloVe

10.5.3 Predatory Identification

Following the predatory conversation detection, we also conduct predatory identification. For all the analysis in this section, we used the results from the best performing feature set and classification algorithm from the previous section. In particular, we used the conversations that were classified as predatory by the Linear SVM classifier on the TF-IDF feature set. For predator identification, we extracted the same type of features and applied the same algorithms as for predatory conversation detection. For training, we extracted all known predatory conversations from the training data and split each conversation into two texts. One was composed by merging all the messages from the sexual predator, while the other text was composed by merging all the messages of the victim. The extracted features were then used to train the various systems. Testing was done by considering all the conversations that were marked as predatory in the previous stage. So some of these messages were, in fact, misclassified non-predatory conversations. For each conversation now, we extracted two texts again in the same way as described for the training data. The features of each text were then classified according to the trained models. The results are presented in the Table 10.5.

When comparing our results to the results from the PAN 2012 competition in [30], we note that these results are slightly lower. The best result was obtained by Villatoro et al. [61] with an $F_{0.5}$-score of 0.935, compared to our best $F_{0.5}$-score of

Feature	Classifier	Accuracy	Precision	Recall	F_1	$F_{0.5}$	F_2
GloVe	LinearSVM	0.864	0.87	0.85	0.860	0.866	0.854
	NonLinSvm	0.859	0.86	0.85	0.855	0.858	0.852
	RF	0.826	0.85	0.81	0.830	0.842	0.818
	NB	0.726	0.80	0.59	0.679	0.747	0.623
TF-IDF	LinearSVM	**0.905**	**0.92**	**0.89**	**0.905**	**0.914**	**0.896**
	NonLinSvm	**0.905**	**0.92**	**0.89**	**0.905**	**0.914**	**0.896**
	RF	0.842	0.84	0.85	0.845	0.842	0.848
	NB	0.905	0.91	0.89	0.900	0.906	0.894
Binary	LinearSVM	0.889	0.90	0.88	0.890	0.896	0.884
	NonLinSvm	0.889	0.90	0.88	0.890	0.896	0.884
	RF	0.854	0.85	0.86	0.855	0.852	0.858
	NB	0.900	0.92	0.88	0.900	0.912	0.888
TC	LinearSVM	0.884	0.89	0.87	0.880	0.886	0.874
	NonLinSvm	0.884	0.89	0.87	0.880	0.886	0.874
	RF	0.851	0.87	0.83	0.850	0.862	0.838
	NB	0.902	0.92	0.88	0.900	0.912	0.888

Table 10.5: Results for predator identification

0.914. The highest F_1-score (0.905) obtained was, however, higher than the one from Villatoro et al. [61], or any other submission to the PAN 2012 competition [30]. It has to be noted that there was a slight difference in the present analysis making the results hard to compare. The goal of the PAN 2012 competition was to identify sexual predators, but most sexual predators had multiple conversations included in the dataset. This would mean that if there were, for example, five conversations of a sexual predator in the dataset, then identifying the sexual predator in at least one conversation would result in a true positive. In our analysis, however, we present a fair comparison by considering all conversations independent of each other. Hence, missing a sexual predator in a particular conversation would count as a false negative, even if that particular predator was detected in another conversation.

10.6 Conclusions and Future Work

In this work, we have analyzed feature spaces for detecting predatory conversation and predatory identification. Specifically, the feature space is analyzed by employing a better preprocessing approach to indicate the superior performance of existing state-of-art algorithms. To systematically demonstrate this, we have employed the PAN 2012 dataset and studied various feature extraction algorithms. In the first part, using the GloVe vector as feature space with SVM has given accuracy of 0.989 and an F_1-score of 0.930 to detect predatory conversations. While

all of the features extraction approaches performed well with the feature selection or feature reduction, the features extracted from the BoW feature spaces did not improve the performance of the models. Also, in the second part, for predatory identification, various feature spaces were investigated, and the best result achieved had an accuracy of 0.905 and an F_1-score of 0.914. In future work, we intend to increase the performance of our approach for predatory identification by extracting stable features that can handle the word embedding for short texts in a realistic scenario. Secondly, we intend to explore the use of deep networks to improve the performance of detecting predators from predatory conversations.

Chapter 11

Article 6: Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles

Borj, Parisa Rezaee, Kiran Raja, and Patrick Bours. "Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles." 2021 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2021.

11.1 Abstract

Securing the safety of the children on online platforms is critical to avoid the mishaps of them being abused for sexual favors, which usually happens through predatory conversations. A number of approaches have been proposed to analyze the content of the messages to identify predatory conversations. However, due to the non-availability of large-scale predatory data, the state-of-the-art works employ a standard dataset that has less than 10% predatory conversations. Dealing with such heavy class imbalance is a challenge to devise reliable predatory detection approaches. We present a new approach for dealing with class imbalance using a hybrid sampling and class re-distribution to obtain an augmented dataset. To further improve the diversity of classifiers and features in the ensembles, we also propose to perturb the data along with augmentation in an iterative manner. Through a set of experiments, we demonstrate an improvement of 3% over the best state-of-the-art approach and results in an F_1-score of 0.99 and an F_β of 0.94 from the proposed approach.

11.2 Introduction

Children are vulnerable due to the new sexual norms caused by advanced technology and increased time spending on online communities where chats with unknown persons are fully possible. The children can thus be targeted by sexual predators by convincing text messages [205]. Detecting and identifying the predatory chats has been a major problem for parents and law enforcement agencies. However, predatory conversation detection is a complex problem as the offenders apply many techniques to avoid disclosure. The predators may not necessarily discuss about sex in the conversations, but apply different strategies and variations in time, type, and intensity to keep the victim interested and eventually exploit them. The process of gaining the trust of victim is usually called grooming [15]. A common challenge in detecting online sexual predators is collecting the data as the chat providers do not make it publicly available, and accessing them requires legal permission. Of the few available datasets like PAN 2012 competition [30], one can observe the common problem encountered in most machine learning problems [30]. The datasets are heavily imbalanced due to normal conversations representing higher proportions than the predatory conversations. In reality, the percentage of sexual predatory data is 0.25% of the total online data that causes many problems for designing an automated machine learning driven detection models [20]. Such composition of dataset makes the predatory detection a challenging problem as handling the imbalanced dataset for the sexual predatory detection is critical.

In this work, we present a new approach for detecting predatory chat detection by providing a new strategy in handling the imbalance to provide a new approach. Specifically, we present an approach which first creates a balanced class distribution by increasing the minor class with a set of augmented and perturbed data. The balanced class distribution is increased until a 50% balance is obtained by simply augmenting and perturbing the data. With the refined class distribution, we create an ensemble of HistogramBoostedGradient classifiers which directly benefit from the augmented and perturbed data in selecting different set of features for creating ensembles. With the set of experimental validation, we evaluate the proposed approach on PAN 2012 [30] dataset where the proposed approach outperforms the existing approaches. The proposed approach results in a precision of 99%, a recall of 99% and a $F_{0.5}$ score of 94% with a gain of 3% over the recent work which reported a recall of 96%. In the rest of this paper, we first present briefly detail the dataset employed and discuss the imbalanced nature of the dataset in Section 11.3. We then list out few related works which have tried to address the imbalanced nature of sexual predatory data for the convenience of the reader in Section 11.4. We present the proposed approach in Section 11.5 followed by the discussion on results in Section 11.6. To the end, we make concluding remarks and list out few

potential future works.

11.3 Database for Sexual Predatory Detection

A chat conversation typically is one of three types of conversations such as (a) a conversation without sexual topics, (b) a conversation between adults on sexual topics, or (c) a conversation between a predator and a minor victim which is considered as a predatory conversation. The PAN 2012 [30] competition dataset deals with the third category and the data contains the conversations between police officers who pretended to be minors and convicted predators extracted from the PJ website (http://www.perverted-justice.com/). The data also contains the ordinary chat without any sexual content extracted from http://www.irclog.org, and sexual conversation between consenting adults from Omegle (www.omegle.com). In addition to the conversation data, the data also consists of a unique conversation ID to distinguish between the conversations. Each message in a conversation further includes an author ID, a timestamp, and the text of the message [28]. In training data, there are 951 predatory conversations and 8477 non-predatory conversations. The test data contains 1697 predatory samples and 19922 non-predatory conversations. More detail about the applied data and the pre-processing method can be found in [30] and [28].

11.3.1 Constraints of Dataset

As with any other type of data investigation, predatory detection requires pertinent data. The amount of predatory data is much lower than the normal chatlogs, making it challenging to find appropriate subset of data. Further, analyzing the data, we note the heavily imbalance in the data where predatory data is less than 0.25% of the total data [20]. Such imbalance leads to sub-optimal classifiers favoring one class over the other resulting either in underfitting or overfitting. When the training data is highly imbalanced, it becomes more critical as the class with fewer samples is severely under-sampled and causes to not capturing the complete information of the given data. If one does not consider the class imbalance problem, the learning techniques can be overwhelmed by the majority class, and the minority class will be easily ignored. An imbalanced classification problem is a problem where the datasets have skewed distributions. It has several characteristics, including class overlapping, small sample size, and small disjuncts [103]. A predatory dataset can suffer from all these characteristics as there are many overlaps and disjuncts between a predatory conversation and a non-predatory one. In addition, the number of predatory conversation samples is much lower than the non-predatory ones. Also, a chat conversation might contain some sentences or topics that are common in both predatory and non-predatory talks.

11.4 Related Works

Earlier research works mainly have used conventional methods for sexual predatory detection disregarding the imbalanced dataset [93, 76, 61, 30, 47, 52, 28]. However, we restrict our focus to few sample works and focus on works that deal with data imbalance problem within predator detection. Cardei et al. [51] tested several techniques for coping with the imbalanced data in sexual predatory detection, such as cost-sensitive technique and sampling techniques including BalanceCascade [104], and CBO - a clustering-based method using k-means [105]. The authors found that the cost-sensitive model where a cost matrix gave a penalty for misclassifying gained the best performance experimenting on PAN 2012 dataset [30]. Their proposed model had two stages where it investigated behavioural features that cover the users' behaviour on the online platform. It considered the ratio of questions, underage expressions, slang words, and the bag of word feature vectors and obtained an $F_{0.5}$ score of 0.95 [51]. Zuo et al. [63] presented an adaptive fuzzy method for artificial neural networks (ANNs) to address the imbalanced data in sexual predatory detection. They used conventional fuzzy inference based on dense rule and fuzzy rule interpolation to handle the imbalanced dataset in the sexual predatory detection problem. Their method was a combination of an adaptive fuzzy inference-based activation function with the artificial neural networks (ANNs) that extracted BoW and TFIDF as feature sets, classified the data sets, and gained an accuracy of 0.766.

11.5 Proposed Approach

The overall pipeline of the proposed approach is illustrated in the Figure 11.1. The proposed approach starts with the preprocessing of the data, followed by feature extraction using Word2Vec [133] and the proposed strategy of learning the ensemble classifiers as detailed below in this section. As the data is heavily noisy, we first preprocess the data to eliminate the irrelevant entries from PAN 2012 dataset. Based on the common properties of the predator victim contacts, we assert that these kinds of conversations have only two authors. We therefore eliminate all the other conversations that involve multi-parties or have only one author. Further, we discard all the conversations with less than seven messages as such conversations contain too little information to be classified as either predatory or non-predatory. Further, as another refinement, we analyze chat messages to eliminate non-English words that did not provide any special meaning or do not follow standard grammar in a remote manner, have many slang and emoticons. To keep the information as much as it is possible, no stemming or lemmatization in the preprocessing of the data was performed. We then extract the features the from the chat logs that cover the word relationships in different contexts with a low dimensional feature

vector. In order to fully exploit the word analogies, we extract features using the Word2Vec embedding model with pre-trained networks with a 300 dimensional vectors. Word2Vec provides distributed representation of the text data which we further use to design the classifier.

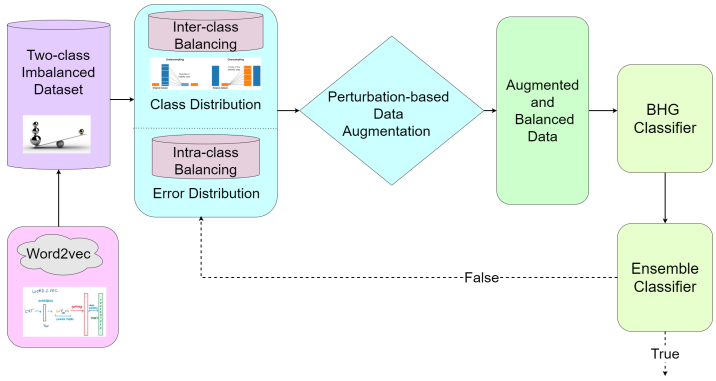

Figure 11.1: Proposed approach for predatory chat detection

11.5.1 Balanced and Augmented Dataset

Given the dataset $\mathcal{D}$ with n classes and m features, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, each class can consist of k samples. When all the classes have equal number of samples, i.e., $k \approx k_{ave}$ one can effectively learn a classifier. However, when the number of samples k_i for a chosen class i is significantly lower than average number of samples from all other classes, the classifier is challenged with skewed data distribution. As it happens, the number of predatory samples is much lower than the number of non-predatory conversation with a sample distribution ratio $1.00 : 8.91$ for predatory to non-predatory samples in our case. Thus, irrespective of the sampling approaches to be used, the minor class will contribute to imbalance for learning a classifier. Thus, we first propose to create an augmented dataset $\mathcal{D}'$ for T number of iterations. For each class C_i in the n classes, we employ two kind of sampling such that class with higher samples is under-sampled and the class with lower samples is over-sampled. In order to achieve this, we simply resort to progressively balanced hybrid sampling using the class distribution. The balanced classes for each iteration is then used to compute the error distribution for the true class distribution and inverse error distribution. Further, as the number of samples in one class can be much higher than the other class, for instance, in our case non-predatory conversations are much higher than the predatory samples, we augment the features in both classes such that the minor class is represented equally with a set of perturbation. The perturbation factor p therefore leads to new samples of the minor class which can be represented as $x' \rightarrow x + \alpha.x^p$ where α is a linear

scaling factor. Thus, the new augmented samples lead to creation of $\mathcal{D}'$. For each of these samples obtained, we obtain new class distribution C' for a given iteration t in total number of iterations T. Using the newly augmented dataset $\mathcal{D}'$ with new class distribution C' with balanced, augmented and perturbed data, we learn a classifier Histogram Gradient Boosted Decision Trees as detailed in the next section. In every iteration t, the class distribution and inverse class distribution is used to balance the samples chosen to learn the classifier.

Algorithm 1: Pseudocode for Proposed Approach

initialization : T iterations, Number of base estimators, Number of bins for HistogramBoostedGradient;
procedure CLASS DISTRIBUTION BALANCE;
t $\leftarrow\in Titerations$
while *K* **do**
    Compute class distributions;
    Compute balanced hybrid sampling;
    Compute the expected class distribution (number of samples from each class);
    Compute the intra-class balanced sampling weights by inversing prediction error distribution;
    Undersample or oversample the features;
end

procedure AUGMENT DATA;
Perturb and augment data;

    procedure CREATE ENSEMBLE;
For each set of augmented data, create classifier - HistogramBoostedGradient;
Fit HistogramBoostedGradient estimator;
Choose features if the loss is less than iteration t-1;

The Algorithm 1 represents the pseudocode of the approach.

11.5.2 Histogram Gradient Boosted Decision Trees

Given the augmented, balanced and perturbed dataset $\mathcal{D}'$ with n samples and m features, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$, the predictions from the boosted decision tree model, $\hat{y}_i$, is defined as a tree-based additive ensemble

model, $\phi(x_i)$, comprising of K additive functions, f_k, defined as:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \;\; f_k \in \mathcal{F}$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}$ is a collection of Classification and Regression Trees, such that $q(x)$ maps each input feature x to one of T leaves in the tree by a weight vector, $w \in \mathbb{R}^T$. Given the function defined above, the Gradient Boosted algorithm minimizes the following regularized objective function:

$$\tilde{\mathcal{L}} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ is the loss function of the ith sample between the prediction $\hat{y}_i$ and the target value y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization component to penalize k^{th} tree in growing additional leaves by λ - a regularization parameter and a weight vector w. We approximate the loss function using a second-order Taylor expansion [206], and we omit the details for the brevity of the paper considering the page limit.

11.5.3 Ensemble Construction

Based on the augmented features selected in each iteration, a classifier is chosen if the loss $l(y_i, \hat{y}_i)$ is the loss function of the ith sample between the prediction $\hat{y}_i$ and the target value y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization component to penalize k^{th} tree in growing additional leaves by λ - a regularization parameter and a weight vector w.

Ref	**Accuracy**	F_1	F_β
Bogdanova et al. [76]	0.97	-	-
Villatoro et al. [61]	0.92	0.87	0.93
Borj & Bours [26]	0.98	0.86	-
Fauzi & Bours [55]	0.95	0.90	0.93
Bours & Kulsrud [50]	-	0.94	0.97
Borj et al. [28]	0.99	0.96	-
Ebrahimi et al. [53]	-	0.80	-
Ebrahimi et al. [52]	0.99	0.77	-
Imbalance based approaches			
Cardei et al. [51]	-	-	0.95
Zuo et al. [63]	0.76	-	-
Proposed Model	**0.99**	**0.99**	**0.94**

Table 11.1: Performance of various approaches against proposed approach. The blocks in gray color indicate the approaches that handle data imbalance and can be directly compared to our proposed approach.

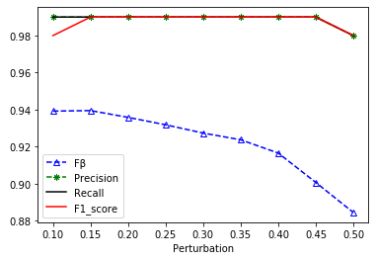

Figure 11.2: Performance variation to perturbation factor in data.

11.6 Experimental Results

For detection of predatory conversation, all the messages of a single conversation were merged into a single text block. Then, we extracted the Word2Vec feature vector for each of the merged texts. The main focus of this analysis is to handle the imbalanced nature of the dataset applying the proposed method. Thus, we select two state-of-the-art approaches which are close to our work to provide a comparison. Specifically, we compare our results against Cardei et al. [51] and Zuo et al. [63] who propose strategies to handle the imbalance in the predatory data. Further, we also compare our results against other state-of-the-art approaches to give a broader comparison. Predatory detection techniques have been evaluated using different metrics such as accuracy, precision, recall, and F_1-score. Further, to avoid many false-positive detection F_β is also recommended as another primary metric for analyzing the performance [30] with $\beta = 0.5$. Table 11.1 demonstrates the obtained results and compares them with the baseline of various works. The proposed approach obtains a gain of 3% over the best benchmark, while it gains more than 23% more accuracy compared to the earlier approach [63] in a similar category of using balancing strategies. Further, we also analyze the effect of perturbation factor in augmenting the dataset, and the obtained accuracy is presented in Figure 11.2. As noted from Figure 11.2, the performance changes slightly when the perturbation factor is increased to more than 20%. Despite the slight drop in performance, one can note the superiority of the proposed approach as compared to the accuracy reported in Table 11.1. Thus, we deduce that the perturbation factor should not be more than 50% to obtain a reliable classification accuracy.

11.7 Conclusion

Predatory conversation detection based on text messages is a crucial problem to avoid exploiting under-aged or minors for sexual favors. Owing to the limited real datasets available, current works employ a standard dataset with less than 10% predatory data leading to a heavy imbalance in the dataset resulting in a classifier that may be sub-optimal. This work has proposed a new approach for handling the imbalanced nature of predatory data by hybrid sampling and class re-distribution to obtain an augmented dataset. Further, to improve the diversity of classifiers and features in the ensembles, this work also proposes to perturb the data along with augmentation in an iterative manner. With the set of experiments on the state-of-the-art dataset, we demonstrate that the proposed approach obtains an improvement over the best state-of-the-art approach by 3% and results in a F_1-score of 0.99 and a F_β of 0.94. Unlike this work, in future works, we also intend to explore different feature extraction approaches to validate the scalability of the proposed approach for predatory detection. Further, this work can also be extended by

generating the predatory data through advanced approaches, including Generative Adversarial Networks.

Chapter 12

Article 7: Detecting Online Grooming By Simple Contrastive Chat Embeddings

Borj, P. R., Raja, K., & Bours, P (2023). Detecting Online Grooming By Simple Contrastive Chat Embeddings, 9th ACM International Workshop on Security and Privacy Analytics (IWSPA 2023) [Accepted].

This paper is not yet published and is therefore not included

Bibliography

[1] Parisa Rezaee Borj, Kiran Raja, and Patrick Bours. Online grooming detection: A comprehensive survey of child exploitation in chat logs. *Knowledge-Based Systems*, page 110039, 2022.

[2] Cornell math explorer's project. http://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/digraphs.html. Accessed: 2019-02-23.

[3] Jamie Bartlett and Louis Reynolds. *The State of the Art 2015: a literature review of social media intelligence capabilities for counter-terrorism*. Demos London, 2015.

[4] Chad MS Steel, Emily Newman, Suzanne O'Rourke, and Ethel Quayle. An integrative review of historical technology and countermeasure usage trends in online child sexual exploitation material offenders. *Forensic Science International: Digital Investigation*, 33:300971, 2020.

[5] Emily A Greene-Colozzi, Georgia M Winters, Brandy Blasko, and Elizabeth L Jeglic. Experiences and perceptions of online sexual solicitation and grooming of minors: a retrospective report. *Journal of child sexual abuse*, 29(7):836–854, 2020.

[6] Ethel Quayle. Prevention, disruption and deterrence of online child sexual exploitation and abuse. In *Era Forum*, volume 21, pages 429–447. Springer, 2020.

[7] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, 34:301022, 2020.

[8] Enrique Guerra and Bryce G Westlake. Detecting child sexual abuse images: traits of child sexual exploitation hosting and displaying websites. *Child Abuse & Neglect*, 122:105336, 2021.

[9] Bryce Garreth Westlake. The past, present, and future of online child sexual exploitation: Summarizing the evolution of production, distribution, and detection. *The Palgrave handbook of international cybercrime and cyberdeviance*, pages 1225–1253, 2020.

[10] Laura Sanchez, Cinthya Grajeda, Ibrahim Baggili, and Cory Hall. A practitioner survey exploring the value of forensic tools, ai, filtering, & safer presentation for investigating child sexual abuse material (csam). *Digital Investigation*, 29:S124–S142, 2019.

[11] Felix Anda, Nhien-An Le-Khac, and Mark Scanlon. Deepuage: improving underage age estimation accuracy to aid csem investigation. *Forensic Science International: Digital Investigation*, 32:300921, 2020.

[12] Pamela Wisniewski. The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience? *IEEE Security & Privacy*, 16(2):86–90, 2018.

[13] Francisco Rangel, Anastasia Giachanou, Bilal Hisham Hasan Ghanem, and Paolo Rosso. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR Workshop Proceedings*, volume 2696, pages 1–18. Sun SITE Central Europe, 2020.

[14] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, pages 1613–0073, 2017.

[15] Samantha Craven, Sarah Brown, and Elizabeth Gilchrist. Sexual grooming of children: Review of literature and theoretical considerations. *Journal of sexual aggression*, 12(3):287–299, 2006.

[16] Nick Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241. IEEE, 2007.

[17] Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka. Characterizing pedophile conversations on the internet using online grooming. *arXiv preprint arXiv:1208.4324*, 2012.

[18] Michael Ashcroft, Lisa Kaati, and Maxime Meyer. A step towards detecting online grooming–identifying adults pretending to be children. In *2015 European Intelligence and Security Informatics Conference*, pages 98–104. IEEE, 2015.

[19] Georgia M Winters and Elizabeth L Jeglic. I knew it all along: The sexual grooming behaviors of child molesters and the hindsight bias. *Journal of child sexual abuse*, 25(1):20–36, 2016.

[20] Matthieu Latapy, Clémence Magnien, and Raphaël Fournier. Quantifying paedophile queries in a large p2p system. In *2011 Proceedings IEEE INFOCOM*, pages 401–405. IEEE, 2011.

[21] Donald A Norman. Memory, knowledge, and the answering of questions. 1972.

[22] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[23] Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*, 2016.

[24] Guoqiang Li, Parisa Rezaee Borj, Loc Bergeron, and Patrick 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)*, pages 1049–1054. IEEE, 2019.

[25] Parisa Rezaee Borj and Patrick Bours. Detecting liars in chats using keystroke dynamics. In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*, pages 1–6, 2019.

[26] Parisa Rezaee Borj and Patrick Bours. Predatory conversation detection. In *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*, pages 1–6. IEEE, 2019.

[27] Parisa Rezaee Borj, Kiran Raja, and Patrick Bours. Detecting sexual predatory chats by perturbed data and balanced ensembles. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021.

[28] Parisa Rezaee Borj, Kiran Raja, and Patrick 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*, pages 1–6. IEEE, 2020.

[29] Parisa Rezaee Borj, Kiran Raja, and Patrick Bours. Detecting online grooming by simple contrastive chat embeddings. In *9th ACM International Workshop on Security and Privacy Analytics (IWSPA 2023) [Submitted and Under-review]*, pages 1–6. ACM, 2023.

[30] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, volume 30, 2012.

[31] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J Wisniewski. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–38, 2021.

[32] Dennis Howitt. *Paedophiles and sexual offences against children.* John Wiley & Sons, 1995.

[33] Rachel O'Connell. A typology of cybersexploitation and online grooming practices. 2003.

[34] Laurence Miller. Sexual offenses against children: Patterns and motives. *Aggression and Violent Behavior*, 18(5):506–519, 2013.

[35] Ming Ming Chiu, Kathryn C Seigfried-Spellar, and Tatiana R Ringenberg. Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words. *Child abuse & neglect*, 81:128–138, 2018.

[36] Loreen N Olson, Joy L Daggs, Barbara L Ellevold, and Teddy KK Rogers. Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251, 2007.

[37] Emily Carmody and Timothy D Grant. Online grooming: moves and strategies. *Language and Law/Linguagem e Direito*, 4(1):103–141, 2017.

[38] April 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.

[39] David Finkelhor. Current information on the scope and nature of child sexual abuse. *The future of children*, pages 31–53, 1994.

[40] Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. Detecting child grooming behaviour patterns on social media. In *International conference on social informatics*, pages 412–427. Springer, 2014.

[41] Fergyanto E Gunawan, Livia Ashianti, Sevenpri Candra, and Benfano Soewito. Detecting online child grooming conversation. In *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, pages 1–6. IEEE, 2016.

[42] Hady Pranoto, Fergyanto E Gunawan, and Benfano Soewito. Logistic models for classifying online grooming conversation. *Procedia Computer Science*, 59:357–365, 2015.

[43] Sana Ali, Hiba Abou Haykal, and Enaam Youssef Mohammed Youssef. Child sexual abuse and the internet—a systematic review. *Human Arenas*, pages 1–18, 2021.

[44] Nur Rafeeqkha Sulaiman and Maheyzah Md Siraj. Classification of online grooming on chat logs using two term weighting schemes. *International Journal of Innovative Computing*, 9(2), 2019.

[45] April Kontostathis, Lynne Edwards, Jen Bayzick, Amanda Leatherman, and Kristina Moore. Comparison of rule-based to human analysis of chat logs. *communication theory*, 8(2), 2009.

[46] Dimitrios Michalopoulos and Ioannis Mavridis. Utilizing document classification for grooming attack recognition. In *2011 IEEE Symposium on Computers and Communications (ISCC)*, pages 864–869. IEEE, 2011.

[47] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. Exploring high-level features for detecting cyberpedophilia. *Computer speech & language*, 28(1):108–120, 2014.

[48] Md Waliur Rahman Miah, John Yearwood, and Sid 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*, pages 157–165, 2011.

[49] Hugo Jair Escalante, Esaú Villatoro-Tello, Antonio Juárez, Manuel Montes, and Luis 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*, pages 46–54, 2013.

[50] Patrick Bours and Halvor Kulsrud. Detection of cyber grooming in online conversation. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[51] Claudia Cardei and Traian Rebedea. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Nat. Lang. Eng.*, 23(4):589–616, 2017.

[52] Mohammadreza Ebrahimi, Ching Y Suen, Olga Ormandjieva, and Adam Krzyzak. Recognizing predatory chat documents using semi-supervised anomaly detection. *Electronic Imaging*, 2016(17):1–9, 2016.

[53] Mohammadreza Ebrahimi, Ching Y Suen, and Olga Ormandjieva. Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, 18:33–49, 2016.

[54] Hugo Jair Escalante, Esaú Villatoro-Tello, Sara E Garza, A Pastor López-Monroy, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111, 2017.

[55] Muhammad Ali Fauzi and Patrick Bours. Ensemble method for sexual predators identification in online chats. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[56] Kanishka Misra, Hemanth Devarapalli, Tatiana R Ringenberg, and Julia Taylor Rayz. Authorship analysis of online predatory conversations using character level convolution neural networks. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 623–628. IEEE, 2019.

[57] Colin Morris. Identifying online sexual predators by svm classification with lexical and behavioral features. *Master of Science Thesis, University Of Toronto, Canada*, 2013.

[58] Fabián Muñoz, Gustavo Isaza, and Luis Castillo. Smartsec4cop: Smart cyber-grooming detection using natural language processing and convolutional neural networks. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 11–20. Springer, 2020.

[59] CH Ngejane, JHP Eloff, TJ Sefara, and VN Marivate. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Science International: Digital Investigation*, 36:301109, 2021.

[60] Javier Parapar, David E Losada, and Alvaro Barreiro. Combining psycholinguistic, content-based and chat-based features to detect predation in chatrooms. *J. UCS*, 20(2):213–239, 2014.

[61] Esaú Villatoro-Tello, Antonio Juárez-González, Hugo Jair Escalante, Manuel Montes-y Gómez, and Luis Villasenor Pineda. A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178, 2012.

[62] Zheming Zuo, Jie Li, Philip Anderson, Longzhi Yang, and Nitin Naik. Grooming detection using fuzzy-rough feature selection and text classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2018.

[63] Zheming Zuo, Jie Li, Bo Wei, Longzhi Yang, Fei Chao, and Nitin 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)*, pages 1–6. IEEE, 2019.

[64] Yun-Gyung Cheong, Alaina K Jensen, Elín Rut Guðnadóttir, Byung-Chull Bae, and Julian Togelius. Detecting predatory behavior in game chats. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):220–232, 2015.

[65] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo 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*, volume 7, 2013.

[66] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[67] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217, 2010.

[68] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[69] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. Technical report, MITRE CORP BEDFORD MA BEDFORD United States, 2011.

[70] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*, 2016:750–784, 2016.

[71] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44, 2011.

[72] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[73] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text-The Hague Then Amsterdam Then Berlin-*, 23(3):321–346, 2003.

[74] Connie Barber and Silvia 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.

[75] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[76] Dasha Bogdanova, Paolo Rosso, and Thamar 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*, pages 110–118, 2012.

[77] Paul Elzinga, Karl Erich Wolff, and Jonas Poelmans. Analyzing chat conversations of pedophiles with temporal relational semantic systems. In *2012 European Intelligence and Security Informatics Conference*, pages 242–249. IEEE, 2012.

[78] Salil P Banerjee and Damon L Woodard. Biometric authentication and identification using keystroke dynamics: A survey. *Journal of Pattern Recognition Research*, 7(1):116–139, 2012.

[79] R Stockton Gaines, William Lisowski, S James Press, and Norman Shapiro. Authentication by keystroke timing: Some preliminary results. Technical report, Rand Corp Santa Monica CA, 1980.

[80] Soumik Mondal, Patrick Bours, Lasse Johansen, Robin Stenvi, and Magnus Øverbø. *Importance of a Versatile Logging Tool for Behavioural Biometrics and Continuous Authentication Research*, chapter 13, pages 282–305. IGI Global, 01 2017.

[81] Romain Giot and Christophe Rosenberger. A new soft biometric approach for keystroke dynamics based on gender recognition. *International Journal of Information Technology and Management*, 11(1-2):35–49, 2012.

[82] Avar Pentel. Predicting age and gender by keystroke dynamics and mouse patterns. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 381–385, 2017.

[83] Ioannis Tsimperidis, Avi Arampatzis, and Alexandros Karakos. Keystroke dynamics features for gender recognition. *Digital Investigation*, 24:4–10, 2018.

[84] Ioannis Tsimperidis, Shahin Rostami, and Vasilios Katos. Age detection through keystroke dynamics from user authentication failures. *International Journal of Digital Crime and Forensics (IJDCF)*, 9(1):1–16, 2017.

[85] Pamela J Black, Melissa Wollis, Michael Woodworth, and Jeffrey T Hancock. A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child abuse & neglect*, 44:140–149, 2015.

[86] Gabriel Dulac-Arnold, Ludovic Denoyer, and Patrick Gallinari. Text classification: A sequential reading approach. In *European Conference on Information Retrieval*, pages 411–423. Springer, 2011.

[87] Fabiola M Villalbos-Castaldi and Ernesto Suaste-Gómez. In the use of the spontaneous pupillary oscillations as a new biometric trait. In *Biometrics and Forensics (IWBF), 2014 International Workshop on*, pages 1–6. IEEE, 2014.

[88] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[89] Michael Fairhurst and Márjory Da Costa-Abreu. Using keystroke dynamics for gender identification in social network environment. 2011.

[90] Syed Zulkarnain Syed Idrus, Estelle Cherrier, Christophe Rosenberger, and Patrick Bours. Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. *Computers & Security*, 45:147–155, 2014.

[91] Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. *University of Massachussetts Amherst, USA*, 2010.

[92] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013.

[93] Vincent Egan, James Hoskinson, and David Shewan. Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex. *Antisocial behavior: Causes, correlations and treatments*, 20(3):273297, 2011.

[94] Tom Buchanan and Monica T Whitty. The online dating romance scam: causes and consequences of victimhood. *Psychology, Crime & Law*, 20(3):261–283, 2014.

[95] Avner Caspi and Paul Gorsky. Online deception: Prevalence, motivation, and emotion. *CyberPsychology & Behavior*, 9(1):54–59, 2006.

[96] Michelle Drouin, Daniel Miller, Shaun MJ Wehle, and Elisa Hernandez. Why do people lie online?"because everyone lies on the internet". *Computers in Human Behavior*, 64:134–142, 2016.

[97] Giuseppe Sartori, Andrea Zangrossi, and Merylin Monaro. Deception detection with behavioral methods: The autobiographical implicit association test, concealed information test–reaction time, mouse dynamics, and keystroke dynamics. In *Detecting Concealed Information and Deception*, pages 215–241. Elsevier, 2018.

[98] David B Buller and Judee K Burgoon. Interpersonal deception theory. *Communication theory*, 6(3):203–242, 1996.

[99] Douglas C Derrick, Thomas O Meservy, Jeffrey L Jenkins, Judee K Burgoon, and Jay F Nunamaker Jr. Detecting deceptive chat-based communication using typing behavior and message cues. *ACM Transactions on Management Information Systems (TMIS)*, 4(2):9, 2013.

[100] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[101] Mihai Varga, George Visu-Petra, Mircea Miclea, and Ioan Buş. The rt-based concealed information test: An overview of current research and future perspectives. *Procedia-Social and Behavioral Sciences*, 127:681–685, 2014.

[102] G Mark Grimes, Jeffrey L Jenkins, and Joseph S Valacich. Assessing credibility by monitoring changes in typing behavior: the keystrokes dynamics deception detection model. In *Hawaii International Conference on System Sciences, Deception Detection Symposium*, 2013.

[103] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[104] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[105] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[106] Batsergelen Myagmar, Jie Li, and Shigetomo Kimura. Transferable high-level representations of bert for cross-domain sentiment classification. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 135–141. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2019.

[107] Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and Ketan Kotecha. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. In *IEEE Access*, volume 9, pages 48364–48404. IEEE, 2021.

[108] Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and Ketan Kotecha. Multi-ideology isis/jihadist white supremacist (miws) dataset for multi-class extremism text classification. In *Data*, volume 6, page 117. MDPI, 2021.

[109] Somya Ranjan Sahoo and Brij B Gupta. Multiple features based approach for automatic fake news detection on social networks using deep learning. In *Applied Soft Computing*, volume 100, page 106983. Elsevier, 2021.

[110] Jitendra Vikram Tembhurne, Md Moin Almin, and Tausif Diwan. Mc-dnn: Fake news detection using multi-channel deep neural networks. In *International Journal on Semantic Web and Information Systems (IJSWIS)*, volume 18, pages 1–20. IGI Global, 2022.

[111] Somya Ranjan Sahoo and Brij B Gupta. Classification of spammer and non-spammer content in online social network using genetic algorithm-based feature selection. In *Enterprise Information Systems*, volume 14, pages 710–736. Taylor & Francis, 2020.

[112] Hadj Ahmed Bouarara. Recurrent neural network (rnn) to analyse mental behaviour in social media. In *International Journal of Software Science and Computational Intelligence (IJSSCI)*, volume 13, pages 1–11. IGI Global, 2021.

[113] Gabriel Weimann. Using the internet for terrorist. *Hypermedia seduction for terrorist recruiting*, 25:47, 2007.

[114] Claudia Peersman, Christian Schulze, Awais Rashid, Margaret Brennan, and Carl Fischer. icop: Live forensics to reveal previously unknown criminal media on p2p networks. *Digital Investigation*, 18:50–64, 2016.

[115] David Bright, Chad Whelan, and Shandon Harris-Hogan. Exploring the hidden social networks of 'lone actor'terrorists. *Crime, Law and Social Change*, 74:491–508, 2020.

[116] Helen C Whittle, Catherine Hamilton-Giachritsis, and Anthony R Beech. Victims' voices: The impact of online grooming and sexual abuse. *Universal Journal of Psychology*, 1(2):59–71, 2013.

[117] Aili Malm, Rebecca Nash, and Ramin Moghadam. Social network analysis and terrorism. *The Handbook of the Criminology of Terrorism*, pages 221–231, 2017.

[118] Pradip Chitrakar, Chengcui Zhang, Gary Warner, and Xinpeng 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)*, pages 493–498. IEEE, 2016.

[119] Katharine Sarikakis and Lisa Winter. Social media users' legal consciousness about privacy. *Social Media+ Society*, 3(1):2056305117695325, 2017.

[120] Brooke Erin Duffy and Ngai Keung Chan. "you never really know who's looking": Imagined surveillance across social media platforms. *New Media & Society*, 21(1):119–138, 2019.

[121] Tatiana Ringenberg, Kanishka Misra, Kathryn C Seigfried-Spellar, and Julia Taylor Rayz. Exploring automatic identification of fantasy-driven and contact-driven sexual solicitors. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pages 532–537. IEEE, 2019.

[122] Miljana Mladenović, Vera Ošmjanski, and Staša Vujičić Stanković. Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. In *ACM Computing Surveys (CSUR)*, volume 54, pages 1–42. ACM New York, NY, USA, 2021.

[123] Cnythia Hombakazi Ngejane, Gugulethu Mabuza-Hocquet, Jan HP Eloff, and Samuel 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)*, pages 1–6. IEEE, 2018.

[124] Kelly M Babchishin, R Karl Hanson, and Chantal A Hermann. The characteristics of online sex offenders: A meta-analysis. *Sexual Abuse*, 23(1):92–123, 2011.

[125] Ming Ming Chiu and Nale Lehmann-Willenbrock. Statistical discourse analysis: Modeling sequences of individual actions during group interactions across time. *Group Dynamics: Theory, Research, and Practice*, 20(3):242, 2016.

[126] Jenny Tam and Craig H Martell. Age detection in chat. In *2009 IEEE International Conference on Semantic Computing*, pages 33–39. IEEE, 2009.

[127] Jane Lin. Automatic author profiling of online chat logs. Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 2007.

[128] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[129] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735, 2008.

[130] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–37, 2008.

[131] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[132] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[133] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[134] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[135] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[136] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke 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)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[137] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[138] Pin Shen Teh, Andrew Beng Jin Teoh, and Shigang Yue. A survey of keystroke dynamics biometrics. *The Scientific World Journal*, 2013, 2013.

[139] Clayton Epp, Michael Lippold, and Regan L Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 715–724, 2011.

[140] Agata Kołakowska. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th international conference on human system interactions (HSI)*, pages 548–555. IEEE, 2013.

[141] Agata Kołakowska. Recognizing emotions on the basis of keystroke dynamics. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 291–297. IEEE, 2015.

[142] Robert Bixler and Sidney 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*, pages 225–234, 2013.

[143] Javier Parapar, David E Losada, and Alvaro Barreiro. A learning-based approach for the identification of sexual predators in chat logs. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178, 2012.

[144] Gunnar Eriksson and Jussi 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.

[145] Nuria Lorenzo-Dus and Cristina Izura. "cause ur special": Understanding trust and complimenting behaviour in online grooming discourse. *Journal of Pragmatics*, 112:68–82, 2017.

[146] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[147] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[148] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[149] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, 2007.

[150] David Hope. Java wordnet similarity library, 2008.

[151] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[152] Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF*, page 2015. sn, 2015.

[153] Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*, 2018.

[154] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, 2009.

[155] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

[156] Mounica Arroju, Aftab Hassan, and Golnoosh Farnadi. Age, gender and personality recognition using tweets in a multilingual setting. In *6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction*, pages 22–31, 2015.

[157] Maite Giménez, Delia Irazú Hernández, and Ferran Pla. Segmenting target audiences: Automatic author profiling using tweets. In *CEUR Workshop Proceedings*, 2015.

[158] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model—notebook for pan at clef 2017. In *CEUR Workshop Proceedings*, volume 1866, 2017.

[159] Braja Gopal Patra, Kumar Gourav Das, and Dipankar Das. Multimodal author profiling for twitter. *Notebook for PAN at CLEF*, 2018.

[160] Robert Veenhoven, Stan Snijders, Daniël van der Hall, and Rik 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)*, volume 2125, 2018.

[161] Roy Khristopher Bayot and Teresa Gonçalves. Multilingual author profiling using lstms. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, 2018.

[162] Jussi Karlgren, Lewis Esposito, Chantal Gratton, and Pentti Kanerva. Authorship profiling without using topical information: Notebook for pan at clef 2018. In *CLEF (Working Notes)*, 2018.

[163] Saman Daneshvar and Diana Inkpen. Gender identification in twitter using n-grams and lsa. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, 2018.

[164] A Pastor López-Monroy, Fabio A González, and Thamar Solorio. Early author profiling on twitter using profile features with multi-resolution. *Expert Systems with Applications*, 140:112909, 2020.

[165] Rafael Dias and Ivandré Paraboni. Cross-domain author gender classification in brazilian portuguese. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1227–1234, 2020.

[166] Lauren Reichart Smith, Kenny D Smith, and Matthew Blazka. Follow me, what's the harm: Considerations of catfishing and utilizing fake online personas on social media. *J. Legal Aspects Sport*, 27:32, 2017.

[167] Aleksei Romanov, Alexander Semenov, Oleksiy Mazhelis, and Jari Veijalainen. Detection of fake profiles in social media-literature review. In *International Conference on Web Information Systems and Technologies*, volume 2, pages 363–369. SCITEPRESS, 2017.

[168] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[169] Liang Li, Weirui Ye, Mingsheng Long, Yateng Tang, Jin Xu, and Jianmin Wang. Simultaneous learning of pivots and representations for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8220–8227, 2020.

[170] Chang Wan, Rong Pan, and Jiefei Li. Bi-weighting domain adaptation for cross-language text classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[171] Joey Zhou, Sinno Pan, Ivor Tsang, and Shen-Shyang Ho. Transfer learning for cross-language text categorization through active correspondences construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[172] Jocelyn Huang, Oleksii Kuchaiev, Patrick O'Neill, Vitaly Lavrukhin, Jason Li, Adriana Flores, Georg Kucsko, and Boris Ginsburg. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*, 2020.

[173] Chongchong Yu, Yunbing Chen, Yueqiao Li, Meng Kang, Shixuan Xu, and Xueer Liu. Cross-language end-to-end speech recognition research based on transfer learning for the low-resource tujia language. *Symmetry*, 11(2):179, 2019.

[174] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388*, 2016.

[175] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[176] Jessica R Blalock and Michael L Bourke. A content analysis of pedophile manuals. *Aggression and Violent Behavior*, page 101482, 2020.

[177] Sarah J Jones, Caoilte Ó Ciardha, and Ian A Elliott. Identifying the coping strategies of nonoffending pedophilic and hebephilic individuals from their online forum posts. *Sexual Abuse*, page 1079063220965953, 2020.

[178] Lyta Penna, Andrew Clark, and George Mohay. A framework for improved adolescent and child safety in mmos. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 33–40. IEEE, 2010.

[179] Katrinna MacFarlane and Violeta Holmes. Agent-mediated information exchange: Child safety online. In *2009 International Conference on Management and Service Science*, pages 1–5. IEEE, 2009.

[180] Dimitrios Michalopoulos, Eustathios Papadopoulos, and Ioannis Mavridis. Artemis: protection from sexual exploitation attacks via sms. In *2012 16th Panhellenic Conference on Informatics*, pages 19–24. IEEE, 2012.

[181] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. In *Psychological bulletin*, volume 129, page 74. American Psychological Association, 2003.

[182] Verónica Pérez-Rosas and Rada Mihalcea. Experiments in open domain deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1120–1125, 2015.

[183] Holly Bentley, Andy Burrows, Laura Clarke, Abbie Gillgan, Jazmin Glen, Maria Hafizi, Fiona Letendrie, Pam Miller, Orla O'Hagan, Priya Patel, Jessica Peppiate, Kate Stanley, Emily Starr, Nikki Vasco, and Janaya Walker. How safe are our children? https://learning.nspcc.org.uk/media/1067/how-safe-are-our-children-2018.pdf, 2018.

[184] Jeffrey John Walczyk, Frank D Igou, Lexie P Dixon, and Talar Tcholakian. Advancing lie detection by inducing cognitive load on liars: a review of relevant theories and techniques guided by lessons from polygraph-based approaches. *Frontiers in psychology*, 4:14, 2013.

[185] Aldert Vrij, Pär Anders Granhag, Samantha Mann, and Sharon Leal. Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science*, 20(1):28–32, 2011.

[186] Ritwik Banerjee, Song Feng, Jun Seok Kang, and Yejin Choi. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1469–1473, 2014.

[187] Bruno Verschuere and Jan De Houwer. Detecting concealed information in less than a second: response latency-based measures. *Memory detection: Theory and application of the Concealed Information Test*, pages 46–62, 2011.

[188] Bram Van Bockstaele, Bruno Verschuere, Thomas Moens, Kristina Suchotzki, Evelyne Debey, and Adriaan Spruyt. Learning to lie: effects of practice on the cognitive cost of lying. *Frontiers in psychology*, 3:526, 2012.

[189] Merylin Monaro, Chiara Galante, Riccardo Spolaor, Qian Qian Li, Luciano Gamberini, Mauro Conti, and Giuseppe Sartori. Covert lie detection using keyboard dynamics. *Scientific reports*, 8(1):1976, 2018.

[190] Reed Abelson. By the water cooler in cyberspace, the talk turns ugly. http://www.nytimes.com/2001/04/29/technology/29HARA.html?searchpv\=site14. 29 April 2001.

[191] A young norwegian boy commited his life. https://www.nrk.no/norge/ung-mann-tok-livet-sitt-etter-a-ha-blitt-utsatt-for-seksuell-utp13943519. Accessed: 2019-02-23.

[192] Avar Pentel. Effect of different feature types on age based classification of short texts. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–7. IEEE, 2015.

[193] Na Cheng, R. Chandramouli, and K.P. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78 – 88, 2011.

[194] Patrick Bours. Continuous keystroke dynamics: A different perspective towards biometric evaluation. *Information Security Technical Report*, 17(1-2):36–43, 2012.

[195] Soumik Mondal and Patrick Bours. Person identification by keystroke dynamics using pairwise user coupling. *IEEE Transactions on Information Forensics and Security*, 12(6):1319–1329, 2017.

[196] Michael C Seto. Pedophilia and sexual offenses against children. *Annual Review of Sex Research*, 15(1):321–361, 2004.

[197] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[198] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765, 2011.

[199] Henk Nijman, Harald Merckelbach, and Maaike Cima. Performance intelligence, sexual offending and psychopathy. *Journal of Sexual Aggression*, 15(3):319–330, 2009.

[200] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[201] Martin Potthast, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. A decade of shared tasks in digital text forensics at pan. In *European Conference on Information Retrieval*, pages 291–300. Springer, 2019.

[202] Michael Frederick McTear, Zoraida Callejas, and David Griol. *The conversational interface*, volume 6. Springer, 2016.

[203] Tianze Shi and Zhiyuan Liu. Linking glove with word2vec. *arXiv preprint arXiv:1411.5595*, 2014.

[204] Alper Kursat Uysal and Serkan Gunal. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112, 2014.

[205] Holly Bentley, Orla O'Hagan, Annie Raff, and Iram Bhatti. How safe are our children. *The most comprehensive overview of child protection in the UK*, 2016.

[206] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[207] World Health Organization et al. What works to prevent violence against children online? 2022.

[208] Jay Morgan, Adeline Paiement, Nuria Lorenzo-Dus, Anina Kinzel, and Matteo Di Cristofaro. Integrating linguistic knowledge into dnns: Application to online grooming detection. 2020.

[209] Nancy Agarwal, Tuğçe Ünlü, Mudasir Ahmad Wani, and Patrick Bours. Predatory conversation detection using transfer learning approach. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 488–499. Springer, 2021.

[210] Matthias Vogt, Ulf Leser, and Alan Akbik. Early detection of sexual predators in chats. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4985–4999, 2021.

[211] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[212] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[213] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[214] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[215] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389, 2022.

[216] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[217] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, 2006.

ISBN 978-82-326-6484-9 (printed ver.)
ISBN 978-82-326-5587-8 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology