HEC MONTRÉAL

Mitigating Online Grooming with Federated Learning

par

Khaoula Chehbouni

HEC Montréal
Directeur de recherche

Sciences de la gestion
(Spécialisation Intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

Aout 2022

Résumé

L’augmentation du temps d’écran et l’isolement dus au confinement ont conduit à une augmentation alarmante du nombre d’incidents d’exploitation sexuelle en ligne. Le pédopiégeage est défini comme étant l’ensemble des stratégies mises en place par les prédateurs pour approcher un enfant à des fins sexuelles. Dans l’industrie et dans le domaine académique, les tentatives pour repérer la prédation sexuelle sur les médias sociaux reposent sur la surveillance des conversations privées des utilisateurs. Nous proposons la première approche de détection préventive du piédopiégeage qui vise à assurer la sécurité des enfants en ligne tout en respectant leur vie privée. Et ce, grâce à un modèle d’apprentissage fédéré qui tient en compte du contexte des conversations, entraîné en respectant la confidentialité différentielle. L’évaluation de notre système de détection sur des données réelles nous indique que sa performance est aussi bonne que celle d’un modèle traditionnel d’apprentissage supervisé. Finalement, nous analysons les compromis nécessaires entre la précision des prédictions, la rapidité de détection et la garantie de protection de confidentialité de notre système.

Mots-clés

Pédopiégeage, Apprentissage fédéré, Apprentissage automatique préservant la confidentialité, Confidentialité différentielle, Détection de risque préventive

Méthodes de recherche

Expérimentation, Exploitation de données, Intelligence artificielle et heuristique

Abstract

The rise in screen time and the isolation brought by the different containment measures implemented during the COVID-19 pandemic have led to an alarming increase in cases of online grooming. Online grooming is defined as all the strategies used by predators to lure children into sexual exploitation. Previous attempts made on the detection of grooming in the industry and academia rely on accessing and monitoring users' private conversations through the training of a model centrally or by sending personal conversations to a global server. We introduce a first, privacy-preserving, cross-device, federated learning framework for the early detection of sexual predators, which aims to ensure a safe online environment for children while respecting their privacy. Empirical evaluation on a real-world dataset indicates that the performance of our framework is as good as the performance of a centrally trained model. Finally, we discuss the necessary trade-offs between the accuracy of a prediction, the speed of the detection and the privacy protection of our framework.

Keywords

Online Grooming, Federated Learning, Privacy-Preserving Machine Learning, Differential Privacy, Early Risk Detection

Research Methods

Experimentation, Data Mining, Artificial Intelligence and Heuristics

Contents

List of Tables

List of Figures

List of Acronyms

DP	Differntial privacy
DP-SGD	Differentially private stochastic gradient descent
eSPD	Early detection of sexual predators
FedAvg	Federated averaging
FedSGD	Federated stochastic gradient descent
FL	Federated learning
IID	Independent and identically distributed
LDA	Latent dirichlet allocation
LSTM	Long short-term memory
PJ	Perverted Justice

Preface

Chapter 1 of this thesis is a reproduction of a research paper co-written with professors Gilles Caporossi (HEC Montreal), Reihaneh Rabbany (McGill University, Mila), Martine De Cock (University of Washington Tacoma) and Golnoosh Farnadi (HEC Montreal, Mila).

Acknowledgements

I would like to thank professor Golnoosh Farnadi, for taking me as a student and for all the amazing opportunities she has given me this past year. Thank you for always being available to answer my questions and for introducing me to the field of responsible AI. I would also like to thank professor Gilles Caporossi, for his expertise and experience and for his contribution to my thesis. This work was also made possible through the invaluable contributions of professors Reihaneh Rabbany and Martine De Cock and their brilliant ideas. It was an honor working with such amazing and inspiring women.

In addition, I would like to aknowledge Mila's AI for humanity, for funding my research and allowing me to quit my job and work full time on this project.

Finally, I would like to thank my friends and family (thank you Najib) for supporting me through my master.

General Introduction

Child sexual abuse is a serious problem that has wide-ranging and lifelong consequences: survivors often suffer from depression, anxiety, low self-confidence, and trust issues throughout their adult life. Suicide attempts, self-harm, panic attacks, eating disorders or alcohol and drug abuse are also reported by victims following the abuse (Jay et al., 2018).

The key to convincing a child to engage in sexual activities is trust. Whether it be online or offline, a potential victim needs to be lured into the relationship. As such, online grooming is defined as "the process whereby an adult seeks to arrange a sexually abusive situation with a minor through the use of cyber-technology" (Lorenzo-Dus et al., 2016).

Every half second around the world, a child goes online for the first time (UNICEF, 2022). And while the internet provides incredible opportunities for creativity, learning, and discovery, it also comes with serious risks. Through social media platforms, chat rooms, or internet games, abusers can lure children into sexual exploitation. By forming an emotional connection and a trusting relationship with their victim, they can convince them into having a sexual conversation, sending sexual content like pictures or videos, and even meet in person.

In 2021 alone, 85 million pictures and videos of child sexual abuse were reported worldwide (European Commission, 2022). And the Internet Watch foundation noted an increase of 64% in reports of confirmed child sexual abuse in 2021 compared to 2020 (European Commission, 2022). In fact, around the world, the COVID pandemic had devastating consequences on children's safety. In Canada, Cybertip [1] – the national tipline for reporting online sexual abuse and exploitation of children – registered a 120% increase in reports of online abuse of children whereas the Royal Canadian Mounted Police's National Child Exploitation Crime Centre reports more than 500 new files a day (Somos, 2022).

Even more alarming, these numbers are restricted to reported cases. Experts agree that the real number of incidences is far greater than what is reported. For example, in the United States, it is estimated that about one in ten children is sexually abused before their eighteenth birthday (Townsend and Rheingold, 2013).

A Model of Deceptive Communication

Thwarting sexual abuse requires insight into the process of operation among predators. To this aim, Olson et al. (2007) presented their luring communication model which details the various strategies used by offline predators to lure children into sexual exploitation.

As seen in Figure 0.1, the *cycle of entrapment*, the model's core phenomenon, is defined as a cyclic approach to luring in which abusers groom, isolate, and, approach their potential victim. These strategies often happen simultaneously, since a variety of techniques are needed to ensure that the victims stay isolated and keep the relationship a secret. The isolation step is essential to the entrapment process: predators need to make sure that their potential victim does not have a strong support system. Thus, children with dysfunctional families, fewer friends, and low self-esteem are often more vulnerable and receptive to their attention.

Olson et al. (2007) define the grooming process as the different strategies predators use to prepare their victim for sexual contact. *Communicative desensitization* and *reframing* are two examples of those strategies. In *communicative desensitization*, predators place themselves in intimate situations with the children to verbally and physically desensitize

[1] https://www.cybertip.ca/en/

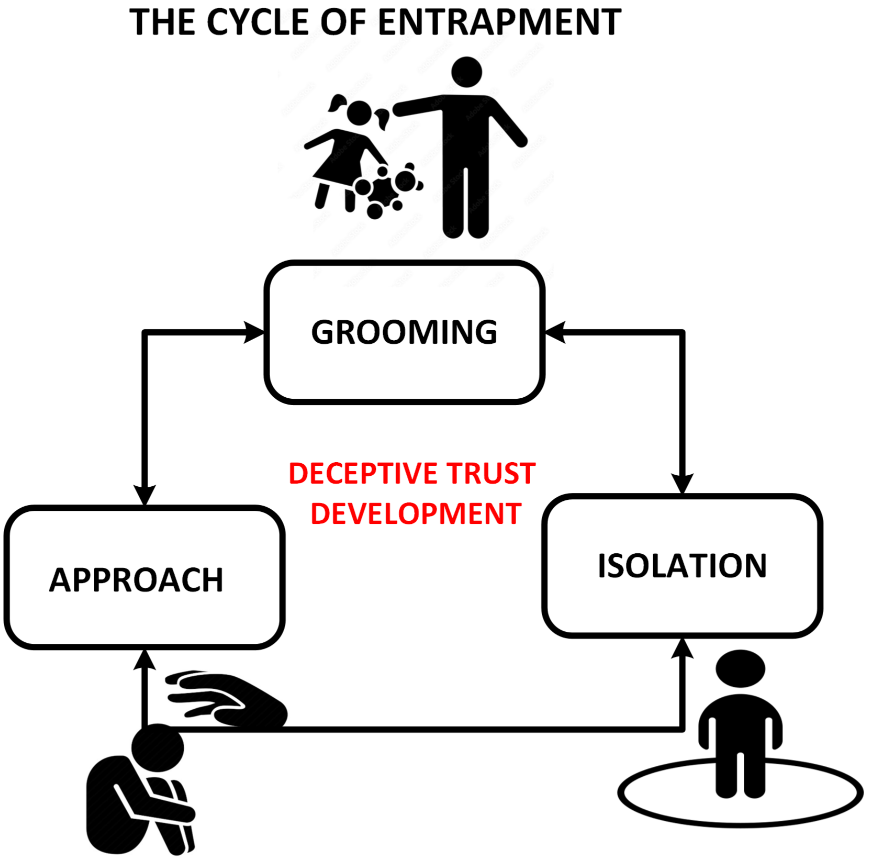


Figure 0.1: Model of luring communication : The cycle of entrapment (Olson et al., 2007)

them to the potential sexual contact. For example, a predator can watch them change their clothes.

Communicative reframing consists of talking about sexual acts in a playful manner: by comparing them to games to play or learning experiences that could be helpful for the child later in life.

Once deceptive trust is established, the groomer can approach the child to see if sexual contact is possible. The approach phase can be viewed as a physical manifestation of how successful the perpetrator has been (Olson et al., 2007). It includes all physical contact or verbal sexual advances. Examples of physical contact may include giving the child a bath or a massage.

Indeed, as noted by journalists Caroline Touzin and Gabrielle Duchaine: "*Le préda-*

teur sexuel n'est plus dans le parc. Il est désormais dans l'écran du cellulaire de votre enfant (Sexual predators are no longer lurking in the neighborhood park: they are behind your child's mobile phone screen)." (Touzin, Caroline and Duchaine, Gabrielle, 2020)

Unfortunately, they are in both. But whereas offline groomers must account for more social stigmas and riskier interactions, online groomers have been shown to be bolder and more direct in their sexual advances (Lorenzo-Dus et al., 2016). Unlike most offline abusers who already know their victims and are part of their entourage, they need to be more creative in gaining the trust of their victims. Online grooming thus requires an intermediate stage of deceptive befriending (Lorenzo-Dus et al., 2016).

To account for these differences, Lorenzo-Dus et al. (2016) adapted Olson et al. (2007) luring communication model. They defined the *entrapment phase:* four interconnected processes of luring communication.

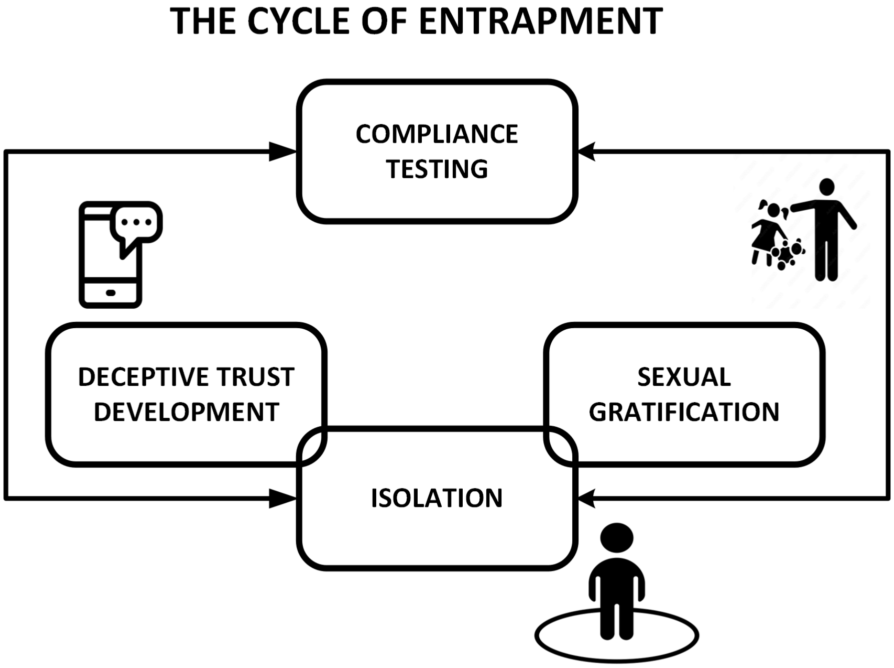


Figure 0.2: The cycle of entrapment for online groomers(Lorenzo-Dus et al., 2016)

As shown in Figure 0.2 *deceptive trust development* and *isolation* corresponds to the phases identified by (Olson et al., 2007). *Deceptive trust development* includes three strategies that also happen in offline settings (exchange of personal information, activities and relationships) and two additional strategies typical of online grooming (praise and sociability). *Praise* refers to all compliments and congratulations made by the groomer to the victim where as *sociability* encompasses the small talk needed to keep the conversation going. Furthermore, they newly identified the *compliance testing* process, comprised of three strategies: *strategic withdrawal*, *role reversal*, and *reverse psychology*. In *strategic withdrawal*, the predator appears to give control to the relationship to their victim. In *role reversal*, the predator act as should be expected from children engaging with adults and in *reverse psychology*, groomers challenge their victim's intent to behave sexually.

By identifying patterns in predators' behavior and mapping their discourse, researchers hope to provide better lexical analysis tools for the detection of sexual predators. Indeed, academic work on the subject has shown that by extracting linguistic and behavioral features from chat logs and using them as input to a classifier, it was possible to identify grooming conversations with high accuracy (Morgan et al., 2021; Inches and Crestani, 2012; McGhee et al., 2011; Pendar, 2007). However, their approach focused on identifying predators after the grooming incurred. Fewer treated the problem from a safety perspective: trying to detect grooming as early as possible (Vogt et al., 2021; Bours and Kulsrud, 2019; López-Monroy et al., 2018).

Trading Privacy for Safety

Despite the concrete need for intervention, one of the major difficulties of implementing safety measures is their toll on the privacy of users. This is because ascertaining the inappropriate nature of a communication requires its interception by a third party.

In the industry, social media platforms like Snapchat, Instagram, and Facebook use advanced technologies and collaborate closely with law enforcement to fight child sexual exploitation online (Picheta, 2019).

A series of attempts to provide more secure systems have been met with outcries from users and privacy experts alike. In 2021, Apple announced two new features for children's protection in iCloud and iMessage (McKinney and Portnoy, 2021). The first feature would compare all pictures when they are uploaded into the cloud with a database of known child pornography. The second feature would scan all messages from minors' accounts for sexually explicit material. Given the intrusive nature of these features, this approach was never implemented by Apple. In addition to breaking end-to-end encryption, and violating users' privacy, experts warned about its dangers and the door it could open to broader abuses (McKinney and Portnoy, 2021).

In May 2022, the European Commission proposed a new regulation to combat child sexual abuse online (European Commission, 2022). The new rules aim to oblige social platform providers to detect, report, and remove child sexual abuse material from their services. The current system, based on voluntary detection and reporting by companies, is not robust enough to prevent the misuse of social platforms for child exploitation. As evidence, in 2020, 95% of the reported cases came from one company only even if the problem is present on all platforms (European Commission, 2022).

This new regulation was however met with wide criticism, deemed even more intrusive than Apple user-side scanning features. Experts argue that scanning users' private messages for child sexual abuse material and grooming behavior is an infringement of privacy and democracy (Vincent, 2022). Indeed, the new regulation involves the possibility for countries to issue "detection orders" to social media platforms. A company that receives such an order is then expected to scan its users' messages for grooming and sexual content. Even if such detection orders are described as being targeted and specified, human rights experts fear the possibility of misuse of such orders for targeting specific individuals: political opponents or minorities (Vincent, 2022).

Contributions

Protecting children from sexual abuse and exploitation should not come at the cost of privacy and additional abuse. In this work, we introduce a privacy-preserving decentralized framework for the early detection of sexual predators. Our model detects grooming as early as possible while ensuring that their personal data is neither monitored nor leaked.

To implement such a framework, we extract a context-aware representation from users' personal conversations using a pre-trained language representation model able to capture the context surrounding a conversation. We then use those features to train a model in a decentralized way using differential privacy to ensure that each users' personal data is not shared or leaked by the resulting model.

Our main contributions are as follow:

- We implement a practical cross-device federated learning framework for the early detection of sexual predators based on users' textual conversation with formal priacy guarantess.
- We propose an extensive evaluation of our framework on a real-world dataset

Thesis Structure

The following sections start by introducing the main machine-learning techniques used for the implementation of our framework. In Chapter 1, we present in its entirety the paper presenting our research: we describe our methods, implementation, and our results. Finally, we conclude by highlighting our contribution and promising results.

Theoretical Framework

We start by presenting the main machine learning techniques used for the implementation of the privacy-preserving framework for the early detection of sexual predators introduced in Chapter 1.

0.1 Bidirectional Encoder Representations from Transformers

In this section, we introduce the Transformer architecture, before presenting BERT, a context-aware language representation model.

0.1.1 Attention is All You Need

In 2017, researchers at Google introduced the Transformer (Vaswani et al., 2017), a novel network architecture more suited for parallelization. With the help of transfer learning and fast computing devices (GPUs and TPUs), the Transformer was able to achieve state-of-the-art results on multiple natural language processing tasks. With parallelization also came scalability: the models were trained in an unsupervised manner on large amounts of raw text to gain a statistical understanding of the language they use.

Vaswani et al. (2017) idea was simple: replace all recurrence and convolution with self-attention mechanisms. Self-attention is an attention mechanism that relates different positions of the same input to compute a representation of the sequence (Vaswani et al., 2017). In other words: “the self-attention mechanism allows the inputs to interact with each other (“self”) and find out who they should pay more attention to (“attention”)” (Raimi, 2019). With self-attention, models can gain a better understanding of the context of a sentence by looking at the relationships between words. Using only self-attention layers, they managed to overcome the main challenges of recurring neural networks: sequential computation and vanishing gradients.

Regarding its architecture, the Transformer follows the encoder-decoder structure used for translation tasks (Cho et al., 2014; Bahdanau et al., 2014; Vinyals et al., 2015). The encoder extract features from an input sequence and returns a contextualized encoding sequence. The decoder models a probability distribution from the given encoded sequence as well as the previously predicted tokens. The encoder is composed of six identical layers: each comprised of two sub-layers: a multi-head self-attention mechanism and a fully connected feed-forward network. The decoder is also composed of six identical layers with the same two sub-layers, but it also has a third sub-layer to perform multi-head attention on the output of the encoder.

0.1.2 Introducing BERT

In 2018, Devlin et al. (2018) introduced a new language representation model called Bidirectional Encoder Representations from Transformers or BERT. The novelty of BERT comes from the fact that, unlike other language representation models like GPT, it can train a deep representation from an unlabeled text using Masked Language Modeling, jointly conditioning on the left and right context in all layers. With its context-aware representation, BERT was able to obtain state-of-the-art results on eleven natural language processing tasks (Devlin et al., 2018).

BERT’s architecture is a multi-layer bidirectional Transformer encoder: the $\text{BERT}_{\text{BASE}}$ model is constituted of 12 layers of transformers blocks with a hidden size of 768 and 12 self-attention heads, with a total of 110 million trainable parameters. Its input representation can handle both single sentences and pairs of sentences, making the model compatible with a variety of downstream tasks.

During pre-training, BERT was trained on unlabeled data on two separate tasks: (1) Prediction of the masked tokens and (2) Next Sentence Prediction. Indeed, to train a deep bidirectional representation, some of the input tokens were randomly masked and the model had to predict them. Furthermore, since language modeling does not directly capture the relationship between sentences, the model was also trained for next-sentence prediction.

This pre-trained deep bidirectional representation alleviates the need for task-specific architectures. As such BERT can be applied to a variety of tasks either using the feature-based approach or by fine-tuning the model. The feature-based approach consists of extracting the contextual embeddings from one or more of the layers without fine-tuning any parameters of the model. The embeddings can then be used as input to a classifier. Whereas the fine-tuning approach relies on fine-tuning all pre-trained parameters on labeled data from the downstream task, usually only for a few epochs.

Vaswani et al. (2017) Transformer architecture brings transfer learning to the forefront, by offering a variety of high-performing models that only need to be fine-tuned rather than trained from scratch. Indeed, BERT's learned understanding of language and context can be re-applied to solve a variety of tasks and achieve state-of-the-art results with few computational resources.

0.2 Federated Learning

This section introduces the background, technical details, and challenges of federated learning.

0.2.1 Background

Meant to take advantage of the unprecedented amounts of data generated each day by edge devices (mobile phones, tablets, wearable devices) and their growing computational power, federated learning (FL) was first introduced in 2017 by McMahan et al. (2017)

as an alternative to privacy-invasive centralized learning. By training statistical models directly on the device, instead of on a global server, federated training ensures that each user's data is never shared. Furthermore, training on local data improves the usability of the resulting models without impacting the user experience or leaking private information (Li et al., 2020).

Federated learning is a new machine learning paradigm, where multiple entities collaborate to train a statistical model under the coordination of a global server (Kairouz et al., 2021). The term was first coined by McMahan et al. (2017) to describe the fact that the "learning task is solved by a loose federation of participating devices (referred to as clients)" and presented as a direct application of the principle of focused collection and data minimization, which aims at protecting consumers' right to privacy (Kairouz et al., 2021). FL only collects relevant information since the updates sent by each client do not contain more information than the raw training data. Furthermore, FL does not retain any personal information, the clients' updates being ephemeral.

Federated learning can take different forms. In *cross-device FL*, millions of entities take part in training. In cross-device training, entities usually share the same features. But the local data distribution varies from one user to another. This type of learning is referred to as *horizontal FL* or *homogeneous FL*. Existing real-life applications of cross-device learning include word prediction, face detection, and voice recognition (Li et al., 2020).

0.2.2 Methodology

The principle behind federated learning is that each entity involved in training receives a model to train using its local data. At the end of the training, the weights are shared with a server, tasked to aggregate them, and update the global model. The process is then repeated until a stopping criterion is attained: a predefined number of rounds or a certain level of performance.

The server operates based on an aggregation algorithm. The baseline algorithm for federated optimization is the federated stochastic gradient descent algorithm (FedSGD). In FedSGD, each client computes the average gradient on its local dataset and the server aggregates these gradients and applies the update. A more used aggregation algorithm is the federated averaging (FedAvg) algorithm, a generalized version of FedSGD. In this setting, multiple iterations of training are made for each client's local update. The average of the weights of the resulting model is then sent to the server for a weighted aggregation that relies on the size of each client's local data. The amount of computation is therefore controlled by three hyperparameters: the fraction of clients selected for training at each round, the number of training passes each client makes over their local data and the local minibatch size used for the client update (McMahan et al., 2017). By increasing the number of local training iterations, McMahan et al. (2017) realized that the resulting models were more stable and able to converge faster, the algorithm acting as a regularizer.

0.2.3 Properties and Challenges of Federated Training

McMahan et al. (2017) defined multiple properties and challenges to federated optimization. They are as follows:

Non-independent and identically distributed data: Whereas centralized training relies on the assumption that the training data's distribution is representative of the whole population, this assumption cannot be made in a federated setting where each user's data has its own distribution. In the case of grooming, this statistical challenge is even more present. In a real-life situation, most users will not interact with predators and thus, will only have access to one label for training.

Imbalanced data: Since each user trains a model on their local data only, the number of examples locally available may vary heavily. For example, for messaging applications, each user has a different sample size. Some people have hundreds of conversations with thousands of messages, whereas others only use the application occasionally and thus have fewer data points to contribute.

Massively distributed: The number of participating devices is expected to be much larger than the number of training examples per client. Indeed, for cross-device training to be efficient, a very large number of clients are expected to take part in training, each contributing with only a small number of examples.

Limited communication and systems heterogeneity: Mobile devices are not always connected or plugged in and often offer fewer computational resources. Furthermore, the storage and computational and communication capabilities of each device differ. This is due to the difference in hardware, network connectivity, and power. Only a fraction of the devices are available at the same time, and dropouts during training are common due to energy or connectivity constraints (McMahan et al., 2017). Federated training must therefore be robust to dropped devices and anticipate heterogeneous hardware.

There are other aspects to consider when implementing a federated system. For example, clients' data change constantly (new messages are sent and others deleted), and there is often a correlation between the clients' availability and their local data distribution (for example, people who speak American English would not be available at the same time as people who speak British or South African English), which could create bias in models. Furthermore, even if the computational costs incurred by federated settings are low, communication is usually very expensive because it relies on connecting a massive number of devices.

Finally, even if federated learning provides significant privacy improvements over centralized training, the baseline model does not provide a formal guarantee of privacy (Kairouz et al., 2021). Indeed, personal data can be inferred from shared updates or from a client's device. The global server could be compromised and training data still be inferred from the deployed model.

0.3 Differential Privacy

This section explains the intuition behind differential privacy.

Whilst state-of-the-art machine learning techniques require an ever increasing volume of data; some models having even been shown to memorize parts of their training data

(Carlini et al., 2022), a growing number of institutions, researchers, and organizations are concerned with privacy challenges. Differential privacy (DP) addresses these concerns by providing a formal and rigorous definition of privacy in machine learning (Dwork, 2008).

With DP comes the promise that everything inferred from a model will be at a population level, not at an individual level. This ensures an individual will not be affected by the use of their private data for training.

DP works by introducing randomness into the training process. The process would be akin to flipping a coin every time an answer to a question as to be provided. Depending on the outcome, the answer will either be the truth or a lie. In the case of a lie, a second coin is flipped to determine which answer (positive or negative) to give (Dwork, 2008). In essence, 50% of the time, the response given is truthful and 50% of the time, the response given is the result of a random coin flip.

For any given answer, it is impossible to know with certainty if it is truthful or not, and that is where the privacy angle comes from. However, with enough answers, it is possible to isolate the randomness and still get accurate models from the aggregated responses.

With such a mechanism, it is now possible to train machine learning model while ensuring the integrity of the training examples.

Chapter 1

Mitigating Online Grooming with Federated Learning

This chapter is a reproduction of a research paper submitted to the 16^{th} International WSDM Conference in August 2022. It was co-written with professor Gilles Caporossi (HEC Montreal), professor Reihaneh Rabbany (McGill University, Mila), professor Martine De Cock (University of Washington Tacoma) and professor Golnoosh Farnadi (HEC Montreal, Mila). The referencing style was changed to match the rest of the thesis.

Abstract

The rise in screen time and the isolation brought by the different containment measures implemented during the COVID-19 pandemic have led to an alarming increase in cases of online grooming. Online grooming is defined as all the strategies used by predators to lure children into sexual exploitation. Previous attempts made on the detection of grooming in the industry and academia rely on accessing and monitoring users' private conversations through the training of a model centrally or by sending personal conversations to a global server. We introduce a first, privacy-preserving, cross-device, federated learning framework for the early detection of sexual predators, which aims to ensure a safe

online environment for children while respecting their privacy. Empirical evaluation on a real-world dataset indicates that the performance of our framework is as good as the performance of a centrally trained model. Finally, we discuss the necessary trade-offs between the accuracy of a prediction, the speed of the detection and the privacy protection of our framework.

1.1 Introduction

With the COVID-19 pandemic, the number of children victim of online grooming has increased substantially: the Canadian Centre for Child Exploitation has recorded an 88% spike in the reported cases of sexual exploitation online (Pawliw, 2021). The unprecedented rise in screen time and isolation brought about by the school closures and lockdowns have left children more vulnerable than ever to online sexual exploitation. In 2021 alone, 85 million pictures and videos of child sexual abuse have been reported worldwide (European Commission, 2022).

Online grooming can be defined as the different strategies used by predators to lure children into sexual relationships. Studies (Olson et al., 2007; Lorenzo-Dus et al., 2016; O'Connell, 2003) have shown that predators have a particular communication style and exhibit common behavior patterns that allow them to approach children, lure them into a trusting relationship, isolate them and desensitize them to the sexual act.

The notable technological advancements these past decades and the proliferation of mobile devices have made children far more accessible. Social media platforms have changed the rules and facilitate unlimited and low-risk access to predators. The direct messages in these platforms are a low-risk tool for luring a child into (online) sexual exploitation. Indeed, while parents can have a tighter hold on their children's interactions in real life, monitoring their online conversations is far more complicated.

In May 2022, the European Commission proposed a new regulation to compel chat apps to scan private user messages for child abuse and exploitation (European Commission, 2022). This new regulation was strongly condemned by privacy experts, who believed that implementing such mechanisms and breaking end-to-end encryption of users' messages could lead to mass surveillance (Vincent, 2022). Other attempts made at adopting child protection features have sparked wide criticism by privacy experts (Snowden, 2021; McKinney and Portnoy, 2021), while existing parental controls, deemed too intrusive, can easily be bypassed. In this paper, we propose a privacy-preserving solution to ensure children's safety in online platforms while protecting their privacy and personal data.

Previous works on the identification of sexual predators have shown that the sexual predators' discourse contains specific indicators that can be exploited for the detection of online grooming. Some researchers focused on finding these linguistic cues by extracting lexical, syntactical, and behavioral features from chat messages (Inches and Crestani, 2012; McMahan et al., 2017). Others have used deep learning techniques to learn useful representations from text (Zambrano et al., 2019; Morgan et al., 2021). Although preventing grooming before any harms occurs is essential to ensure safe access to online social media platforms for children, there are only a few works that treat the grooming detection problem as an early risk detection task (López-Monroy et al., 2018; Vogt et al., 2021). Furthermore, most of the existing work relies on detecting online grooming by monitoring the users' messages and none of the proposed solutions were concerned with ensuring the privacy of the training examples. This represents a major limitation for the applicability of these models in a real-life setting, which is the main focus of this paper.

In this paper we present a novel privacy-preserving decentralized approach to train a context-aware language model for the early detection of sexual predators. To do this, we leverage federated learning, an alternative to centralized machine learning that relies on a global server orchestrating the training of different entities without sharing any raw data. Our key contributions are:

- A practical, cross-device federated learning framework for the early detection of sexual predators based on users' textual conversation with formal differential privacy guarantees.

- An end-to-end implementation of our framework with an extensive evaluation on a real-world dataset.

The remainder of this paper is organized as follows: In Section 1.2 we present the existing work surrounding the task of detecting sexual predators and, more generally, text classification in a decentralized way. In Section 1.3, we introduce the preliminaries of our work, whereas in Section 1.4 we present our framework, and we discuss its implementation details and evaluate it on a real-world dataset in Section 1.5. Finally, we discuss the limitations and opportunities of our framework in Section 1.6 before concluding and presenting possible future works in Section 1.7.

1.2 Related Work

In this section, we review the most relevant works to our proposed approach in three main categories. First, we look at what has been done in the literature for the detection of sexual predators, then we introduce related work on the early detection of sexual predators before presenting existing work on decentralized text classification.

Detection of sexual predators: ChatCoder, a software system designed to identify the different phases of grooming (Olson et al., 2007) was one of the first attempts at classifying predatory conversations. At first they were using dictionaries and a rule-based approach (Edwards and Leatherman, 2009), then they switched to machine learning techniques (McGhee et al., 2011) to classify each message from a conversation extracted from the Perverted Justice (PJ) website[1] into a phase of grooming. They defined four main categories inspired by Olson et al. (2007)'s luring communication model: phase 1: *exchange of personal information*, phase 2: *grooming*, phase 3: *approach*, and *others* for the messages that didn't go into any of the categories. They concluded that their system was more accurate at identifying the non-grooming messages than it was at distinguishing between the different phases of grooming.

[1] perverted-justice.com

In 2012, the international sexual predator identification competition at PAN-12 (Inches and Crestani, 2012) gave greater visibility to the task with the creation of a new annotated dataset for the detection of grooming. Two problems were to be solved: (1) identify the predators among all the users and (2) identify the grooming messages. Most of the participants choose a two-step approach: first identifying the suspicious conversations and then filtering the grooming messages in the flagged conversations. Lexical and behavioral analysis, and pre-filtering of the dataset, were popular techniques used by the participants. Furthermore, most of them chose to not apply any pre-processing technique to the text and considered spelling mistakes, abbreviations, and emojis as lexical features. The winners of the first problem (Villatoro-Tello et al., 2012) used Neural Networks and Support Vector Machines classifiers to identify suspicious conversations on a pre-filtered version of the PAN 12 dataset, whereas the winners of the second problem (Popescu and Grozea, 2012) treated texts as sequences of symbols and used kernel-based learning methods to classify the grooming messages.

Recent work mainly adopted deep learning techniques to solve the task (Zambrano et al., 2019; Morgan et al., 2021). Zambrano et al. (2019) considered the problem as a social engineering attack by first using Latent Dirichlet Allocation (LDA) topic modeling to identify the different phases of grooming before applying a linear classifier to classify the chats. They used a convolutional neural network and a long short-term memory (LSTM) network for the supervised multiclass classification task. Finally, Morgan et al. (2021)'s work integrated linguistic knowledge (Lorenzo-Dus et al., 2016) into the architecture of Deep Neural Networks to improve the classification task.

All these approaches treated the problem from a forensic perspective rather than for prevention. To block harm from occurring, grooming should be detected before a victim is lured into sexual exploitation. Next, we discuss works on the early detection task.

Early risk detection: Another body of work treats the task of identifying sexual predators as an early detection problem (Escalante et al., 2015, 2017; López-Monroy et al., 2018). The main difference between the early detection problem and the standard problem is that while the training phase is similar and the model uses the complete document, during the inference phase, the document is evaluated sequentially at different time steps before being classified.

Escalante et al. (2015) made the first attempt at the early detection of sexual predators by adapting a naïve Bayes classifier for the grooming prediction with partial information. The authors evaluated the performance of their model with different percentages of words from the test set (chunk-by-chunk evaluation). In their follow-up work, they proposed a profile-based representation for the early detection of deception (Escalante et al., 2017). López-Monroy et al. (2018) further extended the profile-based representation by proposing multi-resolution concept representations for the task. They used a chunk-by-chunk evaluation and evaluated their results using different resolutions, i.e. the full document being represented with far more details than a partial read.

In a later work, Bours and Kulsrud (2019) tried to identify how many messages were needed for their classifiers to be able to correctly predict a grooming conversation. They were the first to use full-predatory conversations for the task. They used a TF-IDF representation with a neural network classifier to sequentially classify 10 full PJ conversations. They found that in most cases, 26 to 161 messages of a conversation were sufficient to identify a predator. More recently, Vogt et al. (2021) formally defined the task of early detection of sexual predators (eSPD), moving away from existing work to propose a sliding window evaluation. They also created a new dataset that is better suited for the task. We build on top of this work and use their proposed dataset and their evaluation framework. In this paper, we show that a privacy-preserving framework gives as good a performance as the traditional centralized set-up for the early detection task while ensuring not only the security of children and teenagers online but also their right to privacy.

Our proposed solution relates to the existing work on decentralized text classification, which we will review next.

Federated learning for text classification. The approaches above assume training and deployment of models for grooming detection without concerns for privacy, i.e. while fully disclosing the users' personal messages to a central server for model training. FL, a method for training models in a decentralized fashion at the clients' end, and intermittently aggregating them via a central server, has been proposed as an alternative for natural language processing and text classification tasks (see e.g. (**?**Hilmkil et al., 2021)). While privacy is preserved to some extent in FL because no raw data is disclosed, information about the clients' training data may leak from the gradients or model parameters sent to the central server (Boenisch et al., 2021; Carlini et al., 2022). This information leakage can be mitigated by combining FL with another privacy-enhancing technology such as differential privacy (DP), e.g. by training models with differentially private gradient descent (DP-SGD) (Abadi et al., 2016). Basu et al. (2021) have for instance recently applied FL and DP-SGD for financial text classification. They trained BERT and RoBERTa models on a financial dataset in a centralized setting with differentially private gradient descent (DP-SGD) as well as in a federated setting, showing the necessary trade-off between utility and privacy.

To the best of our knowledge, privacy-preserving early detection of abusive content in a decentralized manner by leveraging both FL and DP-SGD, as we propose in this paper, has not been investigated in the literature.

1.3 Background

In this section, we review several key topics upon which our proposed privacy-preserving early detection of sexual predator framework relies. In our work, we leverage federated learning and differential privacy (DP) to protect the privacy of users, hence we first introduce federated learning and the federated averaging algorithm and then provide a brief overview of the DP-SGD algorithm that we use in our framework.

1.3.1 Federated Learning

Introduced by McMahan et al. (2017) as an alternative to privacy-invasive centralized learning, federated learning (FL) is a machine learning technique that allows multiple entities, called clients, to collaboratively learn a statistical model under the coordination of a central server. The global server orchestrates the training by sampling, at each round, a set of clients to participate in the training (i.e., client selection). Each selected client downloads the current global model, trains it further on its local data and shares a focused update with the server. The server then collects and aggregates all the updates before updating the global model.

Cross-Device and Cross-Silo Federated Learning: There are two main categories of federated learning: cross-device FL and cross-silo FL. Cross-device FL relies on small entities: usually edge devices like smartphones or smart watches. Each client trains a copy of the model on its own personal data. To achieve a good performance, cross-device FL, therefore, requires a very large number of devices participating in the training. In opposition, cross-silo FL involves only a few clients, generally organizations with larger datasets, but everyone of them are expected to participate in the entire training (Huang et al., 2022).

The Federated Averaging Algorithm: The aggregation algorithm used by the global server plays an important role in the federated setting since it defines how the training is going to be orchestrated and how the final model will be computed. The baseline algorithm used for federated optimization is the Federated Stochastic Gradient Descent (FedSGD) algorithm (McMahan et al., 2017): at each round, a fraction of clients is selected, and each client computes the gradient of the loss over its local data. The server then aggregates these gradients and updates the global model. The Federated Averaging algorithm (FedAvg) is a generalization of the FedSGD algorithm where each client is allowed to perform more than one batch update on their local data, and the updated weights rather than the gradients are sent back to be aggregated (McMahan et al., 2017). The server then takes a weighted average of the clients' updates, taking into consideration the amount of data held by each of them to update the global model. Thus, by iterating multiple times on each client, the model is able to converge faster. Furthermore, in addition to lowering communication costs, averaging different models has been shown to act as a regularization technique (McMahan et al., 2017), allowing for more stable models.

1.3.2 Differential Privacy

Whilst FL protects the privacy of the clients by not requiring any raw data to be disclosed, FL in itself does not offer formal privacy guarantees, and the resulting model can leak information about the training data (Carlini et al., 2022). To mitigate such information leakage, FL can be combined with DP (Dwork, 2008) to provide plausible deniability regarding an instance being in a dataset, i.e. offering protection against membership inference attacks.

Formally, DP revolves around the idea of a randomized algorithm – such as an algorithm to train ML models – producing very similar outputs for adjacent inputs. In the context of this paper, two datasets d and d' are considered adjacent if they differ in one record (one labeled instance). A randomized algorithm $M : D \mapsto R$ with domain D and range R is said to be (ε, δ)-differentially private if for any adjacent datasets d and d' and for all subsets of outputs $S \subseteq R$ we have $Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S] + \delta$, where ε is the metric of privacy loss (privacy budget) whereas δ is the probability of data being accidentally leaked. The smaller these values, the stronger the privacy guarantees.

An (ε, δ)-DP randomized algorithm $\mathscr{M}$ is commonly created out of an algorithm $\mathscr{M}^*$ by adding noise that is proportional to the sensitivity of $\mathscr{M}^*$, in which the sensitivity measures the maximum impact a change in the underlying dataset can have on the output of $\mathscr{M}^*$. This technique is used in the differentially private stochastic gradient descent (DP-SGD) algorithm which aims at controlling the influence the training data has on the final model by making the minibatch stochastic optimization process differentially private through clipping and adding noise to the gradients (Abadi et al., 2016). At the end of the training, the overall privacy cost of the mechanism (ε, δ) can be computed from the accumulated costs across all training iterations. Often, a target ε is defined in advance whereas δ should be smaller than the inverse of the size of the training data. We refer to Abadi et al. (2016) for details.

1.4 Methodology

While protecting children from cybercrime is important, the main challenge is the balance between safety and users' privacy. We introduce a privacy-preserving framework for the identification of sexual predators which aims at taking advantage of the growing use of mobile devices by children and teenagers. Our proposed framework consists of, first, training a classifier on the training set (training phase) before evaluating its performance for the early detection task on the test set (inference phase).

1.4.1 Training Phase: eSPD via Federated Learning

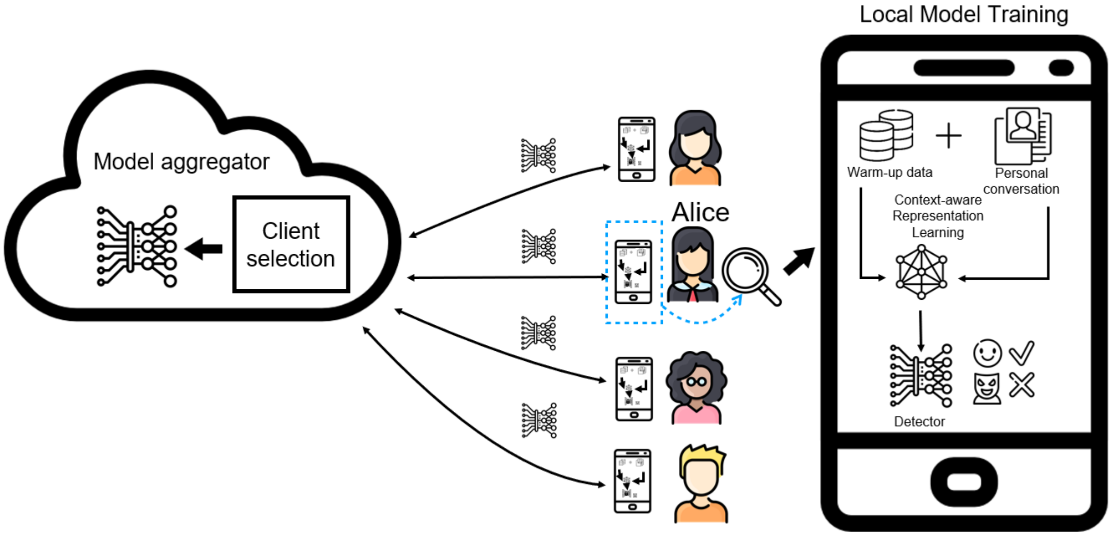


Figure 1.1: Early detection of sexual predators: training phase

We introduce a cross-device federated architecture for the early detection of online grooming: our model is intended to be deployed on each user's cellular device and trained locally on their local data without the need for monitoring them.

Our framework addresses multiple task-specific challenges: (1) training with imbalanced data, (2) training with non-independent and identically distributed (non-IID) data and (3) ensuring that users' personal data are protected during training.

(1) Dealing with imbalanced data. To deal with the problem of imbalanced data – namely very few positive instances – that often comes with early risk detection problems, we implement Errecalde et al. (2017)'s oversampling technique. They considered that the minority class is formed not only by the complete conversation but also by portions of the full conversation at different time steps. Therefore, to account for the sequential nature of the eSPD problem and mitigate the imbalanced nature of the data, we enrich our dataset with chunks of conversations from the minority class, in our case, the conversations with a predator. By giving our system more training examples of the beginning of a conversation with a predator, we are able to gain detection speed.

(2) Training with non-IID data. One of the major challenges of FL is dealing with non-IID data since each client's local data distribution is not representative of the population (Zhu et al., 2021). This statistical challenge is even more prevalent in the context of online grooming since most users are less likely to interact with sexual predators. Thus, the detection of online grooming in a federated setting can be viewed as an extreme case of non-IID data where most users will only have access to one label for training. Indeed, only the victims of online grooming will have access to both grooming and non-grooming conversations.

We use Zhao et al. (2018)'s data-sharing strategy during training in which a small portion of *warm-up data* is distributed to each device in addition to the initial model. The *warm-up data*, which contains public examples from both classes and is balanced, can be seen as a starting point for training, and helps alleviate the statistical challenge. In their paper, Zhao et al. (2018) also suggested sharing a *warm-up model* with each client: a model trained centrally on the warm-up data. We experimented with this strategy but realized that each client did not have enough data to learn from the *warm-up model* and decided to only share a small fraction of *warm-up data* instead.

(3) Protecting users' privacy. Although each client's local data does not leave their device during federated training, it has been shown that it is possible to reconstruct a client's private data using its shared updates (Kairouz et al., 2021), hence a federated architecture by itself does not guarantee privacy. We therefore train each client's model using DP-SGD (see Section 1.3.2), to mitigate leakage of personal information to the server. By clipping the gradient norm of outliers and randomly adding noise during training, we ensure that our model does not memorize any particular information about a single training data point.

Figure 1.1 illustrates the training phase of our framework. A global server selects clients to participate and distributes a model to them; the clients will then further train the model in a privacy-preserving manner on their mobile devices using their own personal data as well as a portion of warm-up data, as we can see in Alice's cellular device.

1.4.2 Inference Phase: Early Detection of Sexual Predators

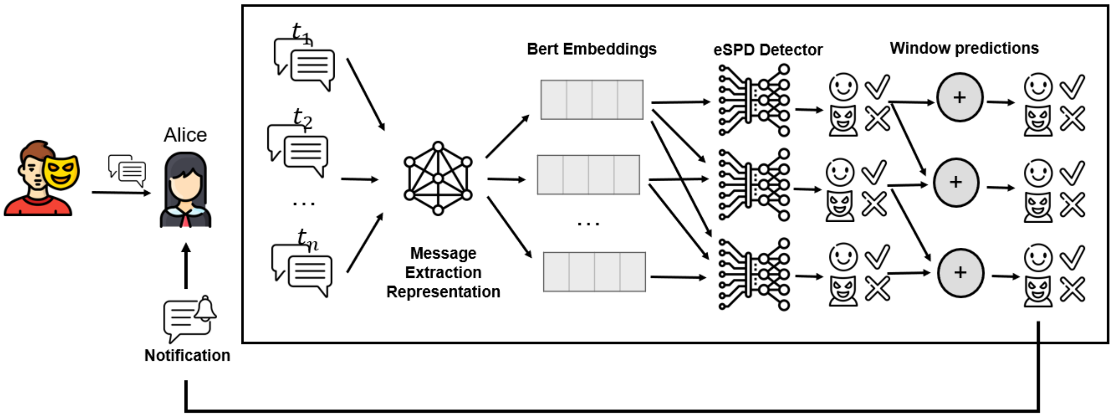


Figure 1.2: Early detection of sexual predators: inference phase

Our work is an extension of the framework proposed by Vogt et al. (2021) for eSPD, i.e. the early risk detection problem (Losada et al., 2020) of sequentially classifying a conversation and detecting early signs of online grooming as soon as possible.

Vogt et al. (2021)'s approach for the inference phase of an eSPD system relies on the use of a sliding window for sequential classification of a conversation. Here, a conversation consists of a sequence of messages $t_1, t_2, \ldots$

For a window of length l, at step s the classifier labels the sequence $t_s, t_{s+1}, \dots, t_{l-1}$, at step $s+1$ the classifier labels the sequence $t_{s+1}, t_{s+2}, \dots, t_l$ etc.

After every window prediction, the system decides whether to raise a warning or not based on the inferred labels of the last 10 window predictions. If a pre-defined threshold – called skepticism level – is reached, a warning is raised and the whole conversation is classified as a grooming conversation. A conversation is only classified as a non-grooming conversation if it is finite and no warning has been raised. Indeed, an eSPD system never classifies a conversation as non-grooming if there are messages left, or if it is still ongoing.

In Figure 1.2, we can see how the different messages received by Alice are analyzed by first being turned into word embeddings and then passed to a classifier given a sliding window for classification. Note that the final prediction is determined based on the previous sequence of predictions and that a warning notification is triggered only when multiple messages are sequentially classified as being grooming messages.

We can envision a system where users will be able to report their own suspicious conversations to the messaging platforms, and will receive a notification if a warning is raised (see Figure 1.3).

1.5 Evaluation

In this section, we show the effectiveness of our proposed approach for the early detection task by performing an empirical evaluation. All our experiments were performed on the *PANC dataset*.

1.5.1 Data

The **PANC dataset** was introduced by Vogt et al. (2021) as a better alternative for the eSPD task. It was created by merging the “negative” (non-grooming) chats from the PAN 12 competition (Inches and Crestani, 2012), sampled from IRC logs[2] and the Omegle

[2] https://www.yoctoproject.org/irc/

forum[3], and the “positive” (grooming) chats from the ChatCoder2 dataset (McGhee et al., 2011): 497 complete conversations extracted from the PJ website and labeled according to the different phases of grooming. They filtered the full grooming conversation and split them into segments to make them comparable to the non-predatory examples and create a corpus better suited for the task of early detection. Despite its numerous limitations, such as the lack of full negative conversations and small differences in formatting between the two classes, we found that the PANC dataset is the most appropriate available data for our task.

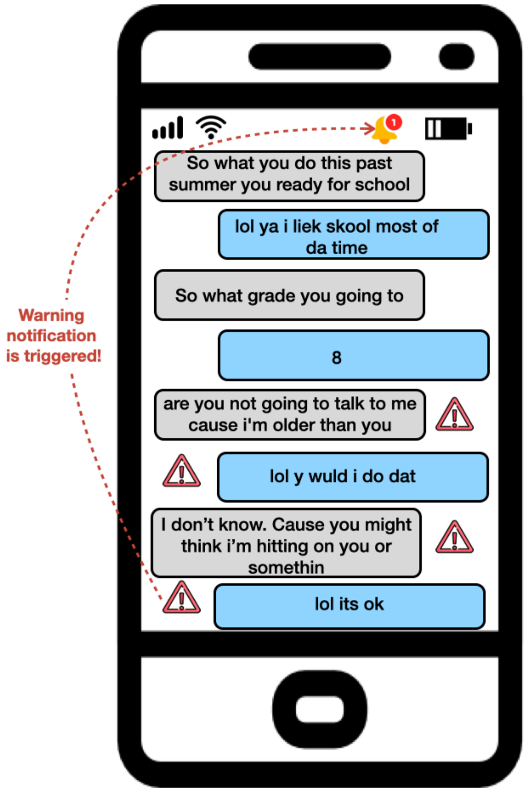
The PANC dataset was split into a training set (60%) and a test set (40%). The training set consists of 1,753 positive segments (representing in total 298 full-length positive chats and 9.06% of the training examples) and 17,598 negative segments, whereas the test set contains 10.84% examples of grooming. Table 1.1 presents the number of the segments and the average number of words they contain for the training set and the test set for each of the classes. We can see that the positive class has in average longer segments. It can be explained by the fact that the full conversations were split into segments of 120 messages, thereby giving longer positive segments since most of them contained 120 messages.

Table 1.1: Statistics about the PANC dataset

	Number of segments		Number of words (mean and std)	
Label	train	test	train	test
0	17598(91%)	11733(89%)	173(±1385)	184(±1529)
1	1753(9%)	1426(11%)	289(±218)	292(±222)

In Figure 1.3, we present a visualization of a synthetic setup based on our proposed framework using a predatory conversation from the PANC dataset. It can take weeks or even months before a warning notification is triggered when a child is being lured by an abuser. Our goal is to minimize the harm by detecting the abuse early and sending a notification to the user. It is up to the user to decide whether to continue the conversation or report the predator. Note that in our framework, both training and inference phases

[3] https://www.omegle.com/



Warning notification is triggered!

So what you do this past summer you ready for school

lol ya i liek skool most of da time

So what grade you going to

8

are you not going to talk to me cause i'm older than you

lol y wuld i do dat

I don't know. Cause you might think i'm hitting on you or somethin

lol its ok

Figure 1.3: Visualization of eSPD in which the risk is detected, warning is raised after passing a threshold, and user is notified as early as possible.

are happening locally and users' personal conversations are never shared with a third-party. Moreover, the global aggregated model from the server can further be tuned and personalized based on users' local data.

1.5.2 Evaluation Metrics

In addition to the established metrics of precision, recall, F1 score, and area under the curve (AUC), we used the latency-weighted F1 score which is introduced by Sadeque et al. (2018) for the early risk detection task. The F-latency metric estimates the trade-off between the speed of detection and the accuracy of the warning by applying a penalty that increases with the warning latency. The warning latency is defined as the number of messages exchanged before a warning is raised (Vogt et al., 2021). The penalty can be computed for each warning latency $l \geq 1$ as follows:

$$\text{penalty}(l) = -1 + \frac{2}{1 + e^{(-p \cdot (l-1))}}$$

where p defines how quickly the penalty should increase. As suggested by Sadeque et al. (2018), p should be set such that the latency penalty is 50% at the median number of messages of a user.

We can then formally define F-latency as:

$$F_{\text{latency}} = F1 \cdot \text{speed}$$

Furthermore, the overall speed of a correct warning is defined as:

$$\text{speed} = 1 - \text{median}\{\text{penalty}(l) \mid l \in \text{latencies}\}$$

where *latencies* corresponds to the list of warning latencies produced by an eSPD system for all grooming chats for which a warning is raised.

Therefore, we only compute the penalty and speed of detection on the positive conversations. It is because the delay needed to detect true positives is a key component of the early risk detection task (Losada et al., 2020; Sadeque et al., 2018). The F-latency metric allows us to adequately evaluate an eSPD system by taking into consideration both the accuracy of the detection and its speed. A higher F-latency score means a better-performing eSPD system.

1.5.3 Experimental Set-Up

In this section, we explain our framework and its implementation details.

Data manipulation: As explained in Section 1.4, we leverage the oversampling technique proposed by Errecalde et al. (2017) to our training data to improve the speed of our system's detection. As such, we have added four additional rows of data to each of the users with a positive label in our training set: the first 10% characters of the full conversation, then 20%, 30%, and 40% of the full conversation. Whereas Errecalde et al. (2017)

selected the above four additional chunks of data to obtain a balanced dataset, we selected the number of augmented data portions with the help of hyperparameter tuning.

Furthermore, to implement the data-sharing strategy, we first split the PANC training set into three portions: 10% of the dataset is randomly selected to create the warm-up data, and the rest is split between a training set (81 %) and a validation set (9%). To ensure that no bias came from the warm-up split, we repeated the process three times and tested our model with every split. We have also experimented with different sizes of warm-up data (1% and 5%) and concluded that a 10% split was better suited for the task (Appendix B). Since neither the test set nor real-life data will be augmented, we removed the additional chunks of data from the validation set. Appendix B shows the new distribution of our dataset and the effect of changing the warm-up data size during training.

Federated set-up: In our cross-device federated framework, each client is randomly selected from the training set. One client corresponds to one user ID, as such since the PANC dataset only has one conversation per user ID, a unique label is associated with the user ID. At initialization, each client receives a random, balanced portion of the warm-up data: ten rows with a “negative” label and ten rows with a “positive” one to complement their own data. Furthermore, if the selected user is a "negative" user, we then select an additional ten "negative" users and combine their data to compensate for the lack of non-predatory examples since they only have a segment of conversation assigned to them. This setup also allows us to simulate a real-life scenario where each client will have multiple conversations to train from. If the selected user has a “positive” label, we leave the data untouched since it is already constituted of multiple segments using the data augmentation technique that we described above.

Choice of the classifier: Although fine-tuning BERT has been shown to give better results for the early detection task, as seen in Appendix B, we use the pre-trained feature-based approach with logistic regression since it is far less computationally expensive and better suited for scaling federated training to a large number of clients. In federated learning, the edge users (clients) are responsible to train the local model on their own devices which can become a bottleneck.

In our framework, each user uses the $\text{BERT}_{\text{BASE}}$ model to create a context-aware representation of their personal conversation by extracting fixed features from the pre-trained model. The [CLS] representation of the last layer is then used as an input for logistic regression with a binary cross entropy loss function. For each user's segment, we, therefore, obtain a 768 length vector.

Implementation: We use Flower (Beutel et al., 2020), an FL framework that facilitates large-scale experiments through its simulation tools, to implement our setup and collaboratively train a logistic regression model with 10,000 clients for 100 rounds. At each round of training, we select 10% of the clients randomly to participate in the training, and the parameters are aggregated with the FedAvg algorithm (McMahan et al., 2017). Surprisingly, the portion of selected clients for training does not have an impact on the performance of the model but heavily influences the computational time. After experimenting with 10%, 25%, 50%, and 100% of selected clients at each round, we have concluded that a 10% fraction fit was enough to achieve a good performance in a reasonable amount of time. Indeed, the computational time is proportional to the number of clients selected for training because in Flower simulation, the training of the clients is sequential, not parallelized since it depends on the number of CPUs available. The optimal number of rounds was determined by following the evolution of the validation loss of different models during training as seen in Appendix D. For the number of clients. We have fine-tuned a model using 100, 1000 and $10,000$ clients and found that the performance usually improves with a larger number of clients. For the results that are presented in Section 1.5.4, we use 10,000 clients to participate in the training phase.

Finally, all of the models were evaluated using a 50-message sliding window and a skepticism level of 5, i.e. 5 of the last 10 predictions had to be positive before a warning was raised. Our federated learning implementation of eSDP will be made publicly available upon publication of the paper.

1.5.4 Empirical Results

We investigate three research questions in our experiments:

RQ1: How is the performance of the eSPD system affected by the federated learning framework?

To address the first research question, we compare the performance of our cross-device approach with three baselines: (1) *Baseline (warm-up data)*: A logistic regression model trained centrally on the warm-up data only, to ensure that our framework is not too biased by the warm-up data distributed to each client. (2) *Centralized LR*: A logistic regression model trained centrally on the training data and the warm-up data; (3) *Cross-Silo FL*: A logistic regression model trained in a federated manner by partitioning the training data and the warm-up data between 5 clients. Both centralized and federated learning models used five-fold cross-validation for hyperparameter tuning whereas the best hyperparameters for the federated models have been chosen using a random search.

Table 1.2: The evaluation results of the *early* online grooming detection task

Model	F1	Recall	Precision	Speed	F-latency
Baseline (warm-up data)	0.50	**0.98**	0.33	0.96	0.48
Centralized eSDP	0.75	0.95	0.62	**0.83**	0.63
Cross-Silo FL eSDP	**0.87**	0.87	**0.87**	0.70	0.61
Cross-Device FL eSDP	0.82	0.85	0.79	0.79	**0.64**
Cross-Device FL+DP-SGD eSDP ($\varepsilon = 1$)	0.76	0.86	0.68	0.81	0.61

In Table 1.2 we can see that the federated frameworks perform better than the centralized models for the early detection task: both the cross-silo and cross-device models show a higher F1 score, compensated with less speed, than the centralized setting. Furthermore, even if the cross-device framework shows slightly better results for the early detection task (with a 64% F-latency score), the cross-silo setup is better at identifying predators with a 87% precision. We think this is due to the large amount of data that is distributed to each client which minimizes the loss of utility that is often attributed to non-IID data. The good performance of the cross-device model for the early detection task can be attributed to the fact that in cross-device training, the minority class is no longer "a minority" since each "positive" user selected for training trains with more grooming examples than non-grooming. Appendix E shows that in a normal evaluation setup, the centralized models still perform better, which is to be expected.

Furthermore, we notice a decrease in utility when making our model $(1, 10^{-5})$-differentially private by training it with DP-SGD. It is interesting to observe that the F-latency is less impacted because of the increase in speed of detection, which can be attributed to the higher recall. Finally, the poor performance of the model trained on the warm-up data confirms that our model is not biased by the data sharing strategy and it is indeed learning from each client's personal data.

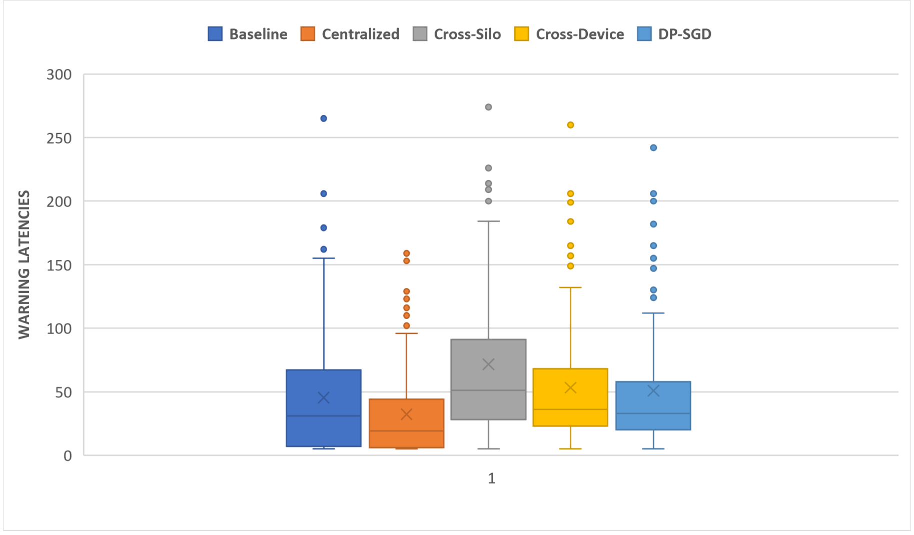


Figure 1.4: Warning latencies for a skepticism level of 5

Figure 1.4 shows the distribution of the warning latencies during the early detection evaluation of our different models.

RQ2: How to reduce the harm of false positives in eSPD?

We look at the impact that a low false positive rate could have on our results. In eSPD, the emphasis is often put on the detection of predators since missing one could cause a lot of harm. Indeed, the F-Latency metric depends on both the F1-score and the speed. And while the F1-score takes into consideration both recall and precision, the speed does not penalize for precision: a model that predicts every conversation as being predatory will have a very high speed. Therefore, we should also evaluate our model to consider the cost of falsely accusing someone. In a real-life setting, we expect an alarm to be raised each time a predator is detected. An innocent could be accused wrongly and it could be even more dangerous if our model is biased towards certain communities like sex workers. Indeed, considering the nature of grooming messages, it is more likely that an eSPD system will be mistaken when confronted with sexual or intimate conversations. Therefore, evaluating an eSPD system should take this into consideration. For this purpose, for each of our models, we identify the classification threshold that is needed to achieve a 1% false positive rate when evaluated on the test set. Using this new threshold, we re-evaluate our models. Table **??** shows that varying the threshold comes with a loss in speed for almost all models except the centralized one and the cross-silo model. It takes a 99% classification threshold for the baseline to achieve a 1% false positive rate showing that it was probably classifying a lot of non-predatory conversations as being predatory conversations. Finally, we notice a decrease in F-latency for our private FL+DP-SGD model, which can be attributed to the large (10%) decrease in speed; which is expected and it is a necessary trade-off to achieve better precision.

Table 1.3: Evaluation results for a 1% FPR

Model	F1	Recall	Precision	Speed	F-latency
Baseline	–	–	–	–	–
Centralized	0.85	0.83	0.88	0.69	0.59
Cross-Device FL	0.83	0.78	**0.89**	**0.73**	**0.61**
Cross-Device FL+DP-SGD ($\varepsilon = 1$)	0.78	0.70	0.88	0.72	0.57

RQ3: How does differential privacy impact the eSPD system?

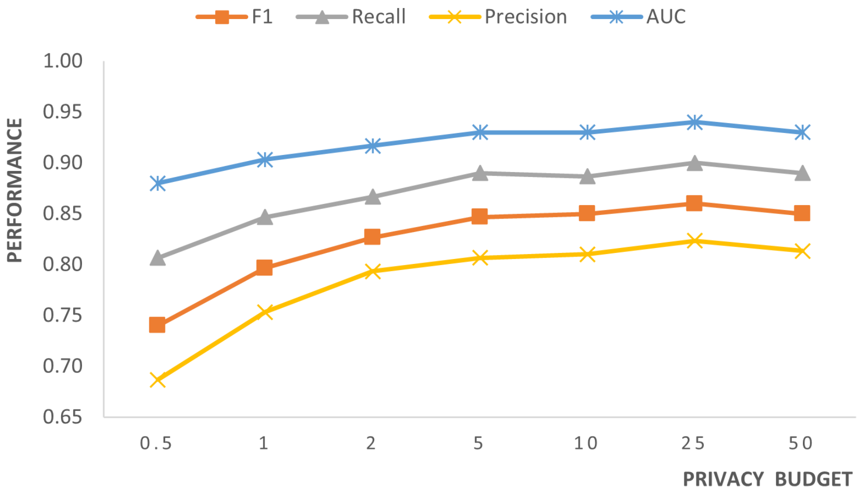


Figure 1.5: Impact of the privacy budget on the performance of our model. All the models were evaluated on the full test set.

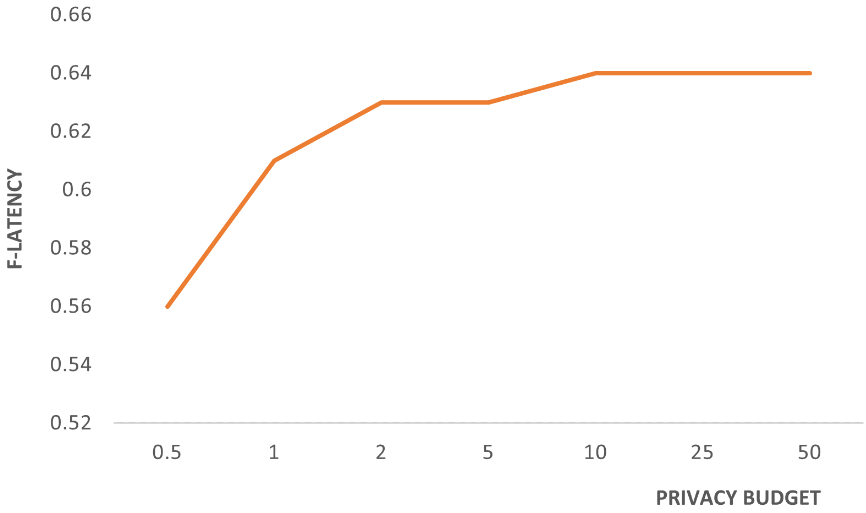


Figure 1.6: Impact of the privacy budget on the early detection performance of our model.

To evaluate the cost of privacy on eSPD systems, we experiment with adding various amounts of noise ε to the training process: every client selected for the training process will train its data with logistic regression with differentially private stochastic gradient descent to ensure that none of its data points is memorized by the model. A random grid search was conducted to test for different hyperparameters: notably, the selected range for the gradient clipping level is $(0.5, 1, 2, 5, 7)$, and we tried $(0.01, 0.05, 0.001, 0.0001)$ for the learning rate, $(8, 16, 32, 100)$ for the batch size , and $(5, 10, 15, 20, 100)$ for the number of local epochs of training. Appendix E shows some of the results obtained during fine-tuning. It is not surprising to observe that the less performing model is the one with the highest privacy constraints: with an ε of 0.50, we notice a drop of 8% of the F-latency score for the most private model as seen in Figure 1.6. However, we notice that there is not much loss of utility when ε is greater than 10 and the loss is negligible when ε is set to 20 or higher. Furthermore, as we can see in Figure 1.5, the precision graph has a steeper slope and therefore seems to be more impacted by the differentially-private training. We can notice a decrease of 12% in the precision score between a model with an ε budget of 5 and a model with an epsilon budget of 0.50 when we evaluate the full test set. Indeed, it has been shown that DP-SGD does not affect the performance of a model equally and that minority classes may be more affected by the training process (Bagdasaryan et al., 2019). In our case, making our model more private may result in a decrease in its ability to detect predators adequately.

1.6 Opportunities and Limitations

In this section, we explore the limitations of our proposed approach and ethical considerations relating to the implementation of such a tool in a real-life setting.

Beyond the privacy issues, a main limitation of the sexual predators' identification task comes from the lack of publicly available labeled and realistic datasets. The different datasets used in the literature take their grooming examples from the PJ website, which are examples of conversations between predators and adults posing as children to catch them. Such chats have been shown to differ from real-life conversations and lack certain aspects of grooming like overt persuasion and sexual extortion (Schneevogt et al., 2018). Indeed, volunteers are often actively trying to get the offenders to be sexually explicit and to arrange an encounter, which is not the case in real-life settings. Furthermore, the non-grooming examples often come from forums and chatrooms where strangers can interact or engage in cyber-sex. Lack of negative examples of trusting and intimate relationships between family members, friends, or partners is an issue of the current datasets which are essential components for a realistic eSPD task.

We hope that the federated architecture we propose in this paper, will give access to a larger range of training examples. Indeed, since each user will be given the option to report abusive content, the conversations flagged as alleged grooming will then be added to the pool of training examples, thus alleviating the lack of realistic and available labeled datasets. Such a system will allow the training examples to be updated regularly, and will consider the growing speed at which language, especially internet slang, evolves.

However, we can imagine that even with such a framework, the labeling will still be an issue since it will rely on users self-reporting cases of grooming. We could think of a preliminary training phase with real data of convicted predators before deploying a pre-trained model to evaluate each user's personal conversation and send a notification where a warning is raised by the eSPD system. Such a model will also alleviate the privacy cost since the first training phase will happen on publicly available data. In this setting, the user will be able to give feedback on the model's prediction. But such a set-up is certainly not ideal, since actual victims of online grooming often trust their abuser and may not realize that they are being manipulated. Notifying a third party, such as a legal guardian or a social worker tasked with monitoring the flagged content, may increase the chances of a case of grooming being reported but will undoubtedly infringe on the privacy of the victim.

Involving law enforcement could also have disastrous consequences. As we have mentioned in subsection 1.5.4, the resulting model could be biased towards certain populations like sex workers, people from the LGBTQI+ community, or people prone to online dating. Evaluating and selecting the best model based on a classification threshold that guarantees a 1% false positive rate can be a first step towards ensuring that the eSPD system does not falsely incriminate. Furthermore, pre-trained language models used to extract a context-aware representation of personal conversations, like BERT, have been shown to reproduce racial and gender biases (Liang et al., 2021). Using such models as a basis for identifying potential suspects to be prosecuted could lead to unanticipated outcomes.

Finally, the literature and datasets used for our experiments concern male predators, both heterosexual and homosexual, that do not know their victims. The lack of data available about female abusers does not allow us to assume that our model is applicable to the detection of female predators.

1.7 Conclusion and Future Directions

In the wake of the new European Commission's regulation (European Commission, 2022), social media companies will be expected to take action to ensure that their underage users are safe from sexual exploitation when using their platforms. Doing so would entail breaking end-to-end encryption and monitoring users' content, which can easily lead to human rights infringements as we have seen recently with the case of the teenager charged for abortion in Nebraska after Meta turned over her personal chat messages to the police (Collier and Burke, 2022). Alternatives to existing privacy-invasive monitoring systems are therefore more pressing since the COVID-19 pandemic has increased the need for children's safety. In this paper, we presented a possible alternative to the existing frameworks for the early detection of sexual predators that will enable a privacy-preserving solution for the detection of online grooming and could pave ways to collect more labeled data.

We presented a first-of-its-kind decentralized framework for the early detection of sexual predators and we showed that the performance of our eSPD system is comparable to the performance of a model trained in a centralized manner while fully protecting users' personal data rights.

Differentially-private optimization approaches that adapt the noise to guarantee privacy have been shown to impact the performance of minority subgroups in federated models (Cummings et al., 2019; Jagielski et al., 2019). We have also noticed that the minority class seemed to be more impacted by the addition of privacy in our framework, hence our eSPD system which leverages DP-SGD algorithm is not able to detect predatory messages as early as the eSPD system without privacy guarantees. Further research is needed to determine the extent to which the issue is raised in a real-world setting. Evaluating the cost of privacy of a model that behaves falsely due to noisy data could be extremely challenging. A proper procedure to label messages as predatory should therefore be clearly defined (as discussed in Section 1.6). Investigating and addressing these challenges are open problems and remain as future directions of this work.

Finally, we believe that our framework can be extended to any early risk detection problem. Future work could explore the use of our framework for the detection of cyberbullying or depression.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Basu, P., Roy, T. S., Naidu, R., and Muftuoglu, Z. (2021). Privacy enabled financial text classification using differential privacy and federated learning. *arXiv preprint arXiv:2110.01643*.

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P. P., and Lane, N. D. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I., and Papernot, N. (2021). When the curious abandon honesty: Federated learning is not private. *arXiv preprint arXiv:2112.02918*.

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

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2022). Quantifying memorization across neural language models.

Collier, K. and Burke, M. (2022). Facebook turned over chat messages between mother and daughter now charged over abortion.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. (2019). On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Edwards, A. and Leatherman, A. (2009). Chatcoder: Toward the tracking and categorization of internet predators. Citeseer.

Errecalde, M. L., Villegas, M. P., Funez, D. G., Ucelay, M. J. G., and Cagnina, L. C. (2017). Temporal variation of terms as concept space for early risk prediction. In *Clef (working notes)*.

Escalante, H. J., Montes-y Gómez, M., Villaseñor-Pineda, L., and Errecalde, M. L. (2015). Early text classification: a naïve solution. *arXiv preprint arXiv:1509.06053*.

Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y Gómez, M., and Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111.

European Commission (2022). Fighting child sexual abuse: Commission proposes new rules to protect children.

Hilmkil, A., Callh, S., Barbieri, M., Sütfeld, L. R., Zec, E. L., and Mogren, O. (2021). Scaling federated learning for fine-tuning of large language models. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–23. Springer.

Huang, C., Huang, J., and Liu, X. (2022). Cross-silo federated learning: Challenges and opportunities.

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

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. (2019). Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

López-Monroy, A. P., González, F. A., Montes-y Gómez, M., Escalante, H. J., and Solorio, T. (2018). Early text classification using multi-resolution concept representations. In *NAACL-HLT*, pages 1216–1225.

Lorenzo-Dus, N., Izura, C., and Pérez-Tattam, R. (2016). Understanding grooming discourse in computer-mediated environments. *Discourse, Context & Media*, 12:40–50.

Losada, D. E., Crestani, F., and Parapar, J. (2020). Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*.

McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., and Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122.

McKinney, I. and Portnoy, E. (2021). Apple's plan to "think different" about encryption opens a backdoor to your private life.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Morgan, J., Paiement, A., Lorenzo-Dus, N., Kinzel, A., and Cristofaro, M. D. (2021). Integrating linguistic knowledge into {dnn}s: Application to online grooming detection.

O'Connell, R. (2003). A typology of cybersexploitation and online grooming practices.

Olson, L. N., Daggs, J. L., Ellevold, B. L., and Rogers, T. K. (2007). Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

Pawliw, B. (2021). Canadian centre for child exploitation reports 88% spike since pandemic.

Popescu, M. and Grozea, C. (2012). Kernel methods and string kernels for authorship analysis. In *CLEF (Online Working Notes/Labs/Workshop)*. Citeseer.

Sadeque, F., Xu, D., and Bethard, S. (2018). Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Schneevogt, D., Chiang, E., and Grant, T. (2018). Do perverted justice chat logs contain examples of overt persuasion and sexual extortion? a research note responding to chiang and grant (2017, 2018). *Language and Law/Linguagem e Direito*, 5(1):97–102.

Snowden, E. (2021). The all-seeing "i": Apple just declared war on your privacy.

Villatoro-Tello, E., Juárez-González, A., Escalante, H. J., Montes-y Gómez, M., and Pineda, L. V. (2012). A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178.

Vincent, J. (2022). New eu rules would require chat apps to scan private messages for child abuse.

Vogt, M., Leser, U., and Akbik, A. (2021). 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.

Zambrano, P., Torres, J., Tello-Oquendo, L., Jácome, R., Benalcázar, M. E., Andrade, R., and Fuertes, W. (2019). Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach. *IEEE Access*, 7:142129–142146.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zhu, H., Xu, J., Liu, S., and Jin, Y. (2021). Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.

General Conclusion

Social media platforms and messaging applications have become the primary form of communication for children and teenagers. And whilst the internet is full of opportunities, it also exposes its more vulnerable users to substantial hazards, online grooming being one of the most pervasive.

To ensure that children are not lured into sexual exploitation, social media platforms must have a robust prevention system in place. Unfortunately, such a system would require having access to personal exchanges, thus breaking end-to-end privacy guarantees such as encryption. This could lead to other misuses of the technology as well as infringe on human rights. To make sure that we are not trading one evil for another, an appropriate eSPD system should therefore ensure that the privacy of the users is preserved.

As part of this research project, we present the first privacy-preserving framework for early detection of sexual predators. Leveraging federated learning, language representation models and differential privacy, we implement a framework for the detection of online grooming which addresses two specific challenges of the task. First, our detection system respects users' privacy while still ensuring children's safety. Second, our detection system enables our model to continuously learn from users' inferred labels through its reporting option.

We were able to achieve high quality results for the task. Our cross-device federated framework achieved a better performance at predicting predators earlier than a traditional centralized setting. As well, our differential-private framework showed competitive results, despite a slight decrease in utility. In addition to implementing such a framework, we demonstrated how the accuracy of a predictor can be impacted by privacy-related decisions, underlining the importance of striking an appropriate balance between the two. This trade-off is of substantial importance in the context of grooming detection since leveling false accusations may have devastating consequences.

Indeed since the detection of grooming often relies on identifying sexual content, we believe that such a system may be more biased against certain individuals: people who use messaging applications with a legitimate suggestive discourse such as sex work or dating. For this reason, we propose to evaluate the eSPD system by setting the classification threshold in such a way that the false positive rate is minimal to account for the high cost of falsely accusing someone.

Furthermore, we warn against the possible racial and gender biases that come with the use of large pre-trained models (Liang et al., 2021). Such a system should therefore never be used directly by law-enforcement agencies at the risk of exacerbating existing social inequalities and persecuting innocents.

In this project, we have shown that safety should not come at the cost of privacy and that less-intrusive monitoring systems for social media applications are possible. We believe that our framework can be generalized to other early risk detection problems on social media like that of cyberbullying or depression and that our setup is model agnostic. Exploring other privacy preserving machine learning techniques like secure aggregation can also be a future direction of our work.

Bibliography

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Basu, P., Roy, T. S., Naidu, R., and Muftuoglu, Z. (2021). Privacy enabled financial text classification using differential privacy and federated learning. *arXiv preprint arXiv:2110.01643*.

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P. P., and Lane, N. D. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I., and Papernot, N. (2021). When the curious abandon honesty: Federated learning is not private. *arXiv preprint arXiv:2112.02918*.

Bours, P. and Kulsrud, H. (2019). Detection of cyber grooming in online conversation.

In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2022). Quantifying memorization across neural language models.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Collier, K. and Burke, M. (2022). Facebook turned over chat messages between mother and daughter now charged over abortion.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. (2019). On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Edwards, A. and Leatherman, A. (2009). Chatcoder: Toward the tracking and categorization of internet predators. Citeseer.

Errecalde, M. L., Villegas, M. P., Funez, D. G., Ucelay, M. J. G., and Cagnina, L. C. (2017). Temporal variation of terms as concept space for early risk prediction. In *Clef (working notes)*.

Escalante, H. J., Montes-y Gómez, M., Villaseñor-Pineda, L., and Errecalde, M. L. (2015). Early text classification: a naïve solution. *arXiv preprint arXiv:1509.06053*.

Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y Gómez, M., and Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111.

European Commission (2022). Fighting child sexual abuse: Commission proposes new rules to protect children.

Hilmkil, A., Callh, S., Barbieri, M., Sütfeld, L. R., Zec, E. L., and Mogren, O. (2021). Scaling federated learning for fine-tuning of large language models. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–23. Springer.

Huang, C., Huang, J., and Liu, X. (2022). Cross-silo federated learning: Challenges and opportunities.

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

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. (2019). Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR.

Jay, A., Evans, S. M., Frank, I., and Sharpling, D. (2018). The effects of child sexual abuse.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

López-Monroy, A. P., González, F. A., Montes-y Gómez, M., Escalante, H. J., and Solorio, T. (2018). Early text classification using multi-resolution concept representations. In *NAACL-HLT*, pages 1216–1225.

Lorenzo-Dus, N., Izura, C., and Pérez-Tattam, R. (2016). Understanding grooming discourse in computer-mediated environments. *Discourse, Context & Media*, 12:40–50.

Losada, D. E., Crestani, F., and Parapar, J. (2020). Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*.

McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., and Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122.

McKinney, I. and Portnoy, E. (2021). Apple's plan to "think different" about encryption opens a backdoor to your private life.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Morgan, J., Paiement, A., Lorenzo-Dus, N., Kinzel, A., and Cristofaro, M. D. (2021). Integrating linguistic knowledge into {dnn}s: Application to online grooming detection.

O'Connell, R. (2003). A typology of cybersexploitation and online grooming practices.

Olson, L. N., Daggs, J. L., Ellevold, B. L., and Rogers, T. K. (2007). Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

Pawliw, B. (2021). Canadian centre for child exploitation reports 88% spike since pandemic.

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

Picheta, R. (2019). Instagram is leading social media platform for child grooming.

Popescu, M. and Grozea, C. (2012). Kernel methods and string kernels for authorship analysis. In *CLEF (Online Working Notes/Labs/Workshop)*. Citeseer.

Raimi, K. (2019). Illustrated:self-attention (corrected).

Sadeque, F., Xu, D., and Bethard, S. (2018). Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Schneevogt, D., Chiang, E., and Grant, T. (2018). Do perverted justice chat logs contain examples of overt persuasion and sexual extortion? a research note responding to chiang and grant (2017, 2018). *Language and Law/Linguagem e Direito*, 5(1):97–102.

Snowden, E. (2021). The all-seeing "i": Apple just declared war on your privacy.

Somos, C. (2022). 'pure evil': How the pandemic has given rise to online child exploitation, livestreamed abuse.

Touzin, Caroline and Duchaine, Gabrielle (2020). L'autre épidémie | le prédateur est dans l'écran.

Townsend, C. and Rheingold, A. A. (2013). Estimating a child sexual abuse prevalence rate for practitioners: studies.

UNICEF (2022). Protecting children online.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Villatoro-Tello, E., Juárez-González, A., Escalante, H. J., Montes-y Gómez, M., and Pineda, L. V. (2012). A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178.

Vincent, J. (2022). New eu rules would require chat apps to scan private messages for child abuse.

Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., and Hinton, G. (2015). Grammar as a foreign language. *Advances in neural information processing systems*, 28.

Vogt, M., Leser, U., and Akbik, A. (2021). 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.

Zambrano, P., Torres, J., Tello-Oquendo, L., Jácome, R., Benalcázar, M. E., Andrade, R., and Fuertes, W. (2019). Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach. *IEEE Access*, 7:142129–142146.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zhu, H., Xu, J., Liu, S., and Jin, Y. (2021). Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.

Appendix A – Effect of the Warm-Up Data

This appendix presents more information about the creation of the warm-up data and the impact its size and distribution have on training.

Effect of The Warm-Up Data on the Speed of Detection

To address the imbalance problem typical of early risk detection, we used Errecalde et al. (2017) temporal variation of terms method to augment our training set. Table 1 shows the new distribution of our dataset for each split: training, warm-up, and validation.

Table 1: Statistics about the new warm-up, train and validation set

	Number of segments			Number of words (mean and std)		
Label	Warm_up	Train	Validation	Warm-up	Train	Validation
0	872(5%)	13513(75%)	1453(8%)	226(±2544)	158(±1448)	136(±133)
1	871(5%)	1276(6%)	93(1%)	300(±310)	312(±427)	262(±194)

We can see that the positive examples now represent 11% of all the data used for training (train set and warm-up set) instead of 7% as in the original PANC training set.

We assumed that giving our classifier more examples from the beginning of the conversations will help it improve its early detection performance. We tested our hypothesis by fine-tuning a centralized model on the augmented training set and compared our performance to Vogt et al. (2021) reported results for the early detection task with $BERT_{BASE}$.

Table 2: Early evaluation of BERT fine-tuned

Model	F1	Precision	Recall	Speed	F-latency
$BERT_{BASE}$ - PANC train	0.81(±0.02)	0.74(±0.05)	0.91(±0.01)	0.63(±0.03)	0.52(±0.03)
$BERT_{BASE}$ - augmented PANC train	0.81(±0.05)	0.69(±0.07)	0.97(±0.005)	0.93(±0.03)	0.75(±0.04)
Vogt et al. (2021) eSPD model	0.89(±0.02)	0.82(±0.04)	0.96(±0.01)	0.91(±0.02)	0.81(±0.03)

In Table 2 we can see that we gain 30% in speed when we fine-tune $BERT_{BASE}$ on our "augmented" training set. It comes however with a loss in precision, which was expected, since there is always a trade-off between utility and speed. We can also notice that Vogt et al. (2021) the eSPD system presents better results. This is because it was trained on the full PANC set and the test set was used as validation, whereas we split the PANC training set between a training set and a validation set, to make sure that our model was capable of generalizing. Nevertheless, our approach seems to have a slightly better speed than theirs (2%).

Effect of the Portion of Data Allocated to the Warm-Up Data

The warm-up data was created using 10% of the training set. In this section, we experiment with different splits in data.

Table 3 presents the results of a federated model trained with 10000 clients on 100 rounds, 10% of them being selected randomly at each round for training. For each user, we add 10 rows of positive warm-up examples and 10 rows of negative warm-up examples for training. And if a negative user is selected, we combine 10 negative conversations.

Table 3: Effect of the portion of data allocated to the warm-up set

Model	WU data	F1	Recall	Precision	AUC
Cross-device FL	1%	0.75	0.72	0.79	0.85
Cross-device FL	5%	0.81	0.8	0.82	0.89
Cross-device FL	10%	0.83	0.85	0.81	0.91

Table 3 shows that better results are obtained whem allocating a bigger portion of the PANC training set to constitute the warm-up set.

Effect of the Size and Distribution of the Warm-Up Data During Training

In this section, we analyze how varying the size and distribution of the warm-up data for each client impacts the training. All models were trained with 10000 clients, but only 10% of them were selected at each round of training. Furthermore, in this part of the analysis, no negative user conversations are combined. One user, therefore, corresponds to one user ID only.

In Table 4, we can see the effect of changing the size of the warm-up data on the training.

Table 4: Effect of the size of the warm-up data

Type	Rounds	WU size	F1	Recall	Precision	AUC
Cross-device FL	100	200	0.81	0.95	0.71	0.95
Cross-device FL	100	2	0.56	0.9	0.4	0.87
Cross-device FL	500	2	0.7	0.79	0.63	0.87
Cross-device FL	500	10	0.74	0.9	0.62	0.92
Cross-device FL	200	20	0.76	0.93	0.64	0.93

When we distribute more rows of training data to each users, the model gets better: with 200 rows of warm-up data per user, we achieve an F1-score of 81% with less rounds of training. With only 2 rows of warm-up data and 10, it takes around 500 rounds to improve the performance of the model. Having more data distributed to each client seems to improve the training. However, since we want to make sure that our model learns from the client's data (as opposed to only the warm-up data), we should make sure that the rows of warm-up data distributed are not bigger than the client's local dataset. We can see that with 20 rows of data, we achieve an acceptable performance with 200 rounds of training, with an F1-score of 76%.

Table 5 shows the effect of varying the distribution of the warm-up data. In Table 4, the warm-up data was balanced. In Table 5 we show how varying the distribution impacts the results: by testing first with 10 rows of warm-up data (9 negative examples and 1 positive example), then 20 rows of data (12 negative examples and 8 positive examples) and finally another 20 rows of data (19 negative examples and 1 positive examples).

Table 5: Effect of the distribution of the warm-up data

Model	WU size	WU dist	F1	Recall	Precision	AUC
Cross-device FL	10	9n-1p	0.43	0.28	0.98	0.64
Cross-device FL	20	12n-8p	0.81	0.88	0.76	0.92
Cross-device FL	20	10n-10p	0.76	0.93	0.64	0.93
Cross-device FL	20	19n-1p	0.3	0.18	0.98	0.59

We can see that varying the distribution of the warm-up dataset has a big impact on training. We observe a poor performance with a highly unbalanced split, whereas a split of 60% of negative examples and 40% of positive examples (12n-8p) seems to have better results than the balanced split. We can assume that this is caused by the fact that negative examples are only constituted of one row of data, while positive examples contain multiple rows (all the different segments that constitute the full conversation to which we have added four additional examples with our data augmentation technique).

Appendix B – Combining Negative Examples

In this section, we evaluate the impact of merging multiple "negative" conversations as being from the same user.

As we have seen earlier, the lack of data for negative users seems to impact the training. Therefore, we try combining multiple conversations to account for the dataset limitation: having only one row of data per negative user. Table 6 shows the results obtained when combining a different number of users. Each model has been trained with 10000 clients, where only 10% are selected at each round of training, for 200 rounds and each user receiving 10 rows of positive warm-up examples and 10 rows of negative warm-up examples.

Table 6: Effect of combining negative users together

Model	Combined	F1	Recall	Precision	AUC
Cross-device FL	10	0.84	0.85	0.82	0.92
Cross-device FL	50	0.82	0.72	0.95	0.86
Cross-device FL	15	0.85	0.83	0.86	0.91
Cross-device FL	0	0.76	0.93	0.64	0.93

We can see that combining negative users improve the performance of our model. For the rest of our experiments, we chose to combine 10 additional users every time a "negative" example is selected as a client because it achieves a good level of utility without merging too many clients.

Appendix C – Effect of the Number of Rounds of Training

In this section, we present the effect the number of rounds of federated training has on utility.

Figure 1 shows the evolution of the validation loss and other metrics during training. The model was trained with 10000 clients, selecting 10% at each round. And each user received 20 rows of balanced warm-up data. Furthermore, for each "negative" client, 10 additional negative examples were merged to constitute the training data. The model was trained for 500 rounds, but the validation loss is stable after a little more than 100 rounds.

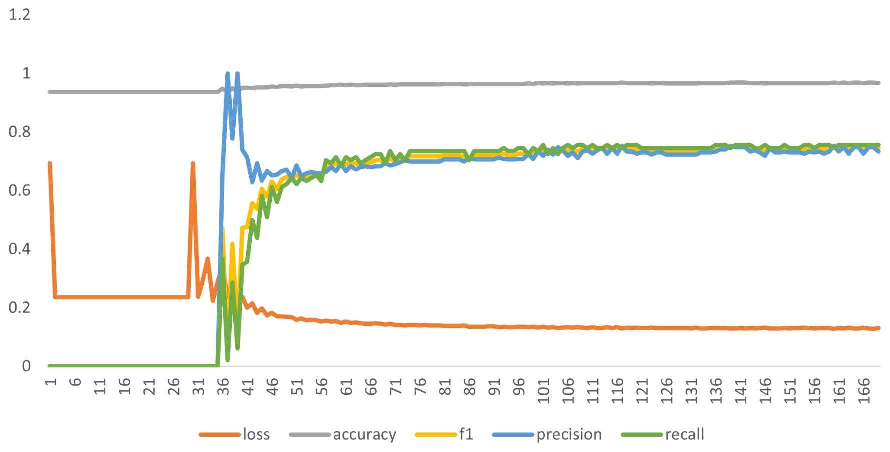


Figure 1: Effect of the number of rounds of training

Therefore, we trained all our models for 100 rounds.

Appendix D – Training with DP-SGD

In this section, we present examples of the results obtained while training our model using differentially-private stochastic gradient descent. We trained all the models with 1000 clients for 100 rounds. Differential private training has a big impact on computational efficiency: the training time is often multiplied by 5 or 10. We, therefore, did our hyperparameters tuning with a sample of the clients. The combinations of hyperparameters showing the best results with 1000 clients were then used to train a model with 10000 clients.

Table 7: Effect of hyperparameter tuning on utility

epsilon	batch	gradient	learning rate	epochs	F1	Recall	Precision	AUC
0.5	100	3	0.001	10	0.39	0.25	0.95	0.62
0.5	100	5	0.0001	10	0	0	0	0.5
0.5	100	2	0.01	15	0.58	0.85	0.44	0.86
1	100	5	0.01	20	0.67	0.87	0.54	0.89
1	100	0.5	0.001	15	0.31	0.22	0.55	0.6
1	100	2	0.001	20	0.54	0.39	0.91	0.69
2	32	7	0.05	15	0.74	0.88	0.64	0.91
10	32	5	0.05	15	0.8	0.94	0.69	0.95
8	100	0.5	0.001	15	0	0	0	0.5

In Table 7 we can see that the utility of the model varies highly when we chose different hyperparameters. The first 3 lines of the table show the results obtained with a privacy budget of $\varepsilon = 0.50$ and we can see that some models have a 0% f1 score. It is probably due to the low learning rate and they will require more rounds of training to achieve a better performance. This table shows the importance of hyperparameter tuning when implementing a model with DP-SGD. Adding noise to training always comes with a drop in utility, but it is possible to gain in performance with the appropriate parameters.

Appendix E – Evaluation of the Full Test Set

In this Appendix, we take a look at the results of the evaluation of our model in a "normal" inference setup: the model is evaluated on the full dataset, without sliding windows.

Table 8: Evaluation of the PANC test set

Model	F1	Recall	Precision	AUC
Baseline (warm-up data)	0.84	**0.95**	0.76	**0.96**
Centralized	**0.92**	0.89	**0.95**	0.94
Cross-Silo FL	0.89	0.84	0.94	0.92
Cross-Device FL	0.85	0.89	0.81	0.93
Cross-Device FL+DP-SGD ($\varepsilon = 1$)	0.79	0.83	0.75	0.90

In a traditional inference setup, we see that the centralized model performs better at identifying sexual predators: here with an F1 score of 92%. This is not surprising since federated learning usually comes with a slight decrease in utility. The baseline model trained on the warm-up data also shows very good results, despite the fact that it was only trained on a small portion (10%) of the training set: with the best AUC score of 96%. However, it also achieves one of the lowest precision scores with an extremely high recall. Furthermore, we can see that the Cross-Silo model shows the closest results to the centralized framework. Indeed, because of the large amount of data distributed to each client, the loss of performance attributed to non-IID data is not significant. We also notice that the drop in utility for the differentially private model is more pronounced when evaluating the full dataset.