The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Enhancing Privacy in the Early Detection of Sexual Predators Through Federated Learning and Differential Privacy

Khaoula Chehbouni[1,2], **Martine De Cock**[3], **Gilles Caporossi**[4], **Afaf Taik** [2, 5]
Reihaneh Rabbany [1, 2], **Golnoosh Farnadi** [1, 2]
[1]McGill University
[2]Mila - Quebec AI Institute
[3]University of Washington Tacoma
[4]HEC Montreal
[5]Université de Montréal
khaoula.chehbouni@mila.quebec

Abstract

The increased screen time and isolation caused by the COVID-19 pandemic have led to a significant surge in cases of online grooming, which is the use of strategies by predators to lure children into sexual exploitation. Previous efforts to detect grooming in industry and academia have involved accessing and monitoring private conversations through centrally-trained models or sending private conversations to a global server. In this work, we implement a privacy-preserving pipeline for the early detection of sexual predators. We leverage federated learning and differential privacy in order to create safer online spaces for children while respecting their privacy. We investigate various privacy-preserving implementations and discuss their benefits and shortcomings. Our extensive evaluation using real-world data proves that privacy and utility can coexist with only a slight reduction in utility.

1 Introduction

With the COVID-19 pandemic, the number of children victim of online grooming (OG) has increased substantially: the Canadian Centre for Child Protection has recorded an 815% spike in the reported cases of sexual exploitation online over the last 5 years (Canada 2024). The unprecedented rise in screen time and isolation brought about by the pandemic 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).

OG can be defined as the different strategies used by predators to lure children into sexual relationships. Studies 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 (Olson et al. 2007; Lorenzo-Dus et al. 2016).

Social media platforms have changed the rules in the past decade, and made children far more accessible to predators. The direct messages in these platforms have become 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) and compromise not only users' privacy but also their online safety: online scams, cyberattacks or social engineering attacks being more likely to happen once encryption is lifted (Radauskas, Gintaras 2023). As a result, the e-Privacy derogation was implemented to allow companies to detect and remove child sexual abuse material online on a voluntary basis only. In April 2024, the Parliament decided to prolong the exemption until April 2026 (European Parliament 2024). In the meantime, it is crucial for social media providers to devise more effective strategies for safeguarding children while also respecting the privacy of their users since previous attempts made at adopting child protection features, such as Apple's and Facebook's scanning tools, have sparked wide criticism by privacy experts (Snowden 2021; Hill 2012). More recently, Google made the news when a father was the subject of an investigation for child sexual abuse and exploitation after sending naked pictures of his toddler to the pediatrician (Hill 2022).

Previous work in communication studies has shown that the sexual predators' discourse contains specific indicators that can be leveraged for the detection of online grooming (Olson et al. 2007; Lorenzo-Dus et al. 2016). Some researchers focused on finding these linguistic cues by extracting lexical, syntactical, and behavioral features from chat messages (Inches and Crestani 2012; Mcghee et al. 2011). Others have used deep learning techniques to learn useful representations from text (Zambrano et al. 2019; Morgan et al. 2021). All, however, have been limited by the lack of realistic and publicly-available training datasets for the task. Although preventing grooming before any harm 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), i.e. recognizing grooming while it is happening and intervention is possible, as opposed to detection afterward. 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 work, we explore different privacy-preserving frameworks for the early detection of sexual predators (eSPD) in ongoing conversations. Our proposed solution for this application is the use of Federated Learning (FL) (McMahan et al. 2017a), an alternative to centralized machine learning (ML) that relies on a global server orchestrating the training of models with data originating from different entities without requiring them to share any raw data. This approach holds the potential to address the two primary challenges facing the application: obtaining access to real-world data for training while preserving privacy, and increasing the availability of labeled data through the implementation of a reporting mechanism to identify predators. However, it is important to note that while FL operates by keeping client data local to their devices during the training process, the federated architecture alone does not guarantee complete privacy protection. Studies have demonstrated that it is feasible to reconstruct a client's sensitive information from the shared updates they contribute during the process (Kairouz et al. 2019). We therefore enhance our solution with differential privacy (DP) to provide formal privacy guarantees (Dwork 2008). Our work presents the following contributions: (1) different practical and privacy-preserving frameworks for eSPD; (2) a comprehensive implementation of the proposed frameworks, including a thorough examination of various DP algorithms to enhance privacy, and (3) an in-depth evaluation of the framework using a real-world dataset.

2 Related Work

In this section, we review the most relevant works to our proposed approach.

Detection of sexual predators. A competition organized at PAN-12 [1] attracted attention to the task of identifying sexual predators with the creation of a new annotated dataset for the detection of online grooming in messages (Inches and Crestani 2012). The used techniques varied among the winners, as Villatoro-Tello et al. (2012) used neural networks and SVMs, and Popescu and Grozea (2012) used kernel-based learning methods. In the following years, the problem was approached using representation learning techniques (Liu et al. 2017; Zambrano et al. 2019; Agarwal et al. 2021; Morgan et al. 2021; Rezaee Borj, Raja, and Bours 2023). For instance, Liu et al. (2017) approached the problem as a sentiment analysis task and employed recurrent neural networks with long short-term memory cells to enhance accuracy. Rezaee Borj, Raja, and Bours (2023) for their part showed how a simple contrastive learning framework for feature extraction at a sentence-level achieved state-of-the-art results. Whereas Agarwal et al. (2021) used BERT to extract context-aware representations from text, similar to our approach in this work. However, it is worth noting that these studies approached the problem from a forensic perspective rather than a preventative one.

Early detection of sexual predators. To block harm from occurring, grooming should be detected before a victim is lured. Escalante et al. (Escalante et al. 2016) made the first attempt at the early detection of sexual predators by adapting a naive Bayes classifier for grooming prediction with partial information. The authors evaluated the performance of their model with different percentages of words from the test set in a chunk-by-chunk evaluation framework that was later extended using profile-based representation (López-Monroy et al. 2018). More recently, Vogt et al. (2021) formally defined the eSPD task, moving away from existing work to propose a sliding window evaluation, and creating a new dataset that is better suited for the task. The previously mentioned studies on the detection of grooming have assumed that training and deployment of models could be performed without considering privacy implications. This involves the full disclosure of users' private messages to a central server for model training. Our work extends the study by Vogt et al. (2021) and utilizes their evaluation framework and dataset to demonstrate how eSPD can be adapted in a distributed setting while ensuring formal privacy guarantees.

Privacy-preserving text classification. Various privacy-enhancing technologies for text classification have been proposed in the literature. Solutions based on cryptographic techniques such as secure multiparty computation (MPC) or homomorphic encryption enable privacy-preserving *inference*, i.e. classification of a message held by one party with a model held by another party without requiring each of the parties to disclose their data to anyone in plain text (Reich et al. 2019; Adams, Melanson, and De Cock 2021; Resende et al. 2022; Lee et al. 2022). However, while MPC protocols provide *input privacy*, i.e. the data holders do not disclose their inputs in plain text to anyone, the computed result – such as the trained model – is the same as one would obtain without the use of cryptography; in other words, MPC by itself does not provide *output privacy*. This is problematic as information about the training data can leak from a model and from inferences made with it (Fredrikson, Jha, and Ristenpart 2015; Tramèr et al. 2016; Song, Ristenpart, and Shmatikov 2017; Carlini et al. 2019).In this paper, we use a combination of FL and DP to provide both input *and* output privacy. While the use of FL and DP has received a considerable amount of attention in the natural language processing (NLP) literature (see e.g. (Brown et al. 2022; Klymenko, Meisenbacher, and Matthes 2022) and references therein), only a few authors have looked into the use of FL for supervised learning tasks of text classification. In particular, FL an DP were combined in applications such as sentence intent classification Zhu et al. (2020) and financial text classification Basu et al. (2021). More recently, Shetty et al.

---

[1] https://pan.webis.de/clef12/pan12-web/index.html

(2023) and Samee et al. (2023) leveraged FL for privacy-preserving detection of cyberbullying: (Shetty et al. 2023) introduced FedBully, a framework leveraging FL, sentence-encoders and secure aggregation for privacy-preserving detection of cyber-bullying whereas (Samee et al. 2023) looked at how FL with DP, word embeddings and emotional features could be paired for cyberbullying detection. We focus instead on evaluating how different DP implementations can affect a framework for the early detection of abusive content, and the challenges of such implementations in a real-life scenario.

3 Background

In our work, we leverage FL and DP to protect the privacy of users. Hence, we first introduce FL and then provide a brief overview of the different DP algorithms we implement.

Federated learning. Introduced by McMahan et al. (2017a) as an alternative to privacy-invasive centralized learning, FL is an ML 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. 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. The aggregation algorithm used by the global server plays an important role in the federated setting since it defines how the training is orchestrated and how the final model will be computed. In this work, we use the federated averaging algorithm (FedAvg) (McMahan et al. 2017a).

Differential privacy. Whilst FL protects the privacy of the clients by preventing the sharing of raw data, FL in itself does not offer formal privacy guarantees, and the resulting model can leak information about the training data (Thakkar et al. 2021; Carlini et al. 2022). To mitigate such information leakage, FL can be combined with DP (Dwork 2008) to provide plausible deniability regarding the existence of an instance or a user in a dataset, i.e. offering protection against membership inference attacks (Shokri et al. 2017).

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

DP can be implemented at the *instance-level*: two datasets are considered adjacent if they only differ in one labeled instance, or at the *user-level*: two datasets are considered adjacent if they only differ in one user's complete data. According to the above definition, $\mathcal{M}$ is then (ϵ, δ)-DP if the probability that $\mathcal{M}$ generates a specific model from the data is very similar to the probability of generating that model if a particular instance (resp. user) had been left out of the data. The latter implies that what the model has memorized about individual instances (resp. users) is negligible.

DP is a widely used privacy-preserving technique in the field of machine learning. DP provides formal privacy guarantees for sensitive data by adding controlled noise to the data, thereby obscuring individual contributions.There are two main types of DP: local DP and global DP. Local DP operates on individual data points, whereas global DP operates on aggregate data. And whereas global DP relies on the notion of adjacent inputs, local DP does not. As such, a randomized algorithm $\mathcal{M}$ is said to satisfy ϵ-Local DP if and only if for any pair of input values v and v' in the domain of $\mathcal{M}$ and for any output $y \in \mathcal{Y}$ we have $\mathbb{P}[\mathcal{M}(v) = y] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(v') = y]$, where $\mathbb{P}[\cdot]$ is the probability.

DP can be applied in various stages of the machine learning pipeline: either during pre-training by adding noise to the raw data, or during training by injecting noise into gradients or weights, as further outlined below.

Metric DP. Local differential privacy (LDP) allows each user to perturb their own data, providing strong privacy guarantees, typically at a cost in utility. Indeed, in the context of language, applying LDP means that all word embeddings are exchangeable without accounting for meaning, which may hurt the utility of the downstream task. To tackle this challenge, Feyisetan et al. (2020) adapt a relaxed version of DP, called metric DP, for natural language processing. A randomized mechanism $\mathcal{M} : \mathcal{D} \mapsto \mathcal{R}$ is said to be $\eta d_{\mathcal{X}}$-private if for any x and $x' \in \mathcal{D}$ the distribution over outputs of $\mathcal{M}(x)$ and $\mathcal{M}(x')$ is bounded by: $\mathbb{P}[\mathcal{M}(x) = y] \leq e^{\eta d(x,x')} \cdot \mathbb{P}[\mathcal{M}(x') = y]$ for all $y \in \mathcal{R}$, where d is a distance function $d : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}_{+}$. Qu et al. (2021) adopt this definition for sequence representation privatization: the input to the randomized mechanism $\mathcal{M}$ is a token embedding, or in our case, the [CLS] representation produced by BERT. As such, metric DP can be achieved for the Euclidean distance by adding calibrated noise to the sequence representations of the training data (Qu et al. 2021).

The DP-SGD Algorithm. An (ϵ, δ)-DP randomized algorithm $\mathcal{M}$ is commonly created out of an algorithm $\mathcal{M}^*$ by adding noise that is proportional to the sensitivity of $\mathcal{M}^*$, in which the sensitivity measures the maximum impact a change in the underlying dataset from d to an adjacent dataset d' can have on the output of $\mathcal{M}^*$. DP-SGD is a commonly used technique for training deep neural networks with DP guarantees. DP-SGD is designed to limit the impact that the training data has on the final model by making the mini-batch stochastic optimization process differentially private through gradient clipping and adding noise (Abadi et al. 2016).

The DP-FedAvg Algorithm. McMahan et al. (2017b) introduce a DP version of the FedAvg algorithm to offer user-level privacy. Since the FedAvg algorithm already groups multiple SGD updates together and operates at the user-level, it is possible to extend the DP-SGD algorithm to provide formal privacy guarantees to federated training. In order to achieve (ϵ, δ)-DP, each client clips its own update, and after each round of federated training, the server adds noise to the aggregated updates. The overall privacy cost relies on the number of clients sampled at each round of training, on the number of federated rounds of training, and on a hyperparameter that controls the scale of the noise added to the update: the noise multiplier.

4 Method

Ensuring the safety of children from cybercrime is crucial, but it should not compromise the privacy of social media users. As such, we introduce a privacy-preserving framework for the identification of sexual predators. This novel approach takes advantage of the rising prevalence of mobile devices among children and teenagers. Our code as well as additional resources, including the Appendix, are available at: https://github.com/khaoulachehbouni/fl-espd.

4.1 eSPD Training via Federated Learning

One of the significant challenges in FL is addressing non independent and identically distributed (IID) data, as the local data distribution of each client is not representative of the overall population (Zhu et al. 2021). This challenge becomes even more pronounced in the context of OG, where most users are unlikely to interact with sexual predators. Thus, the detection of OG 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. To address this issue, we implement a 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 (Zhao et al. 2018). 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.

4.2 eSPD Training with Differential Privacy

In this work, we explore how FL and DP can be used to offer formal privacy guarantees in different ways. Figure 5 in Appendix B illustrates our different privacy-preserving frameworks.

(1) Adding noise to the raw data. Each client adds noise directly to the token representations of their data to achieve sequence representation privatization. The perturbed embeddings are then used for training, thus ensuring no access to the raw data. In the rest of this paper, we refer to this implementation as *FL with metric DP*.

(2) Adding noise during training. Each client train their model using DP-SGD to mitigate leakage of personal information to the server. At each round of local training, the clients clip the gradient norm of outliers in their data and adds calibrated noise to the gradients to ensure that the shared updates provide instance-level privacy protections. In this case, the raw data is accessed during local model training, but never sent to a global server. In the rest of this paper, we refer to this implementation as *FL with DP-SGD*.

(3) Adding noise to the updates. We train our federated model using the DP-FedAvg algorithm. After each round of training, the server adds noise to the aggregated updates sent by the clients. This approach provides user-level protections but relies on the hypothesis that the server can be trusted since it has access to non-private updates. In the rest of this paper, we refer to this implementation as *FL with DP-FedAvg*.

5 Evaluation

In this section, we present an empirical evaluation of our different frameworks for the early detection of sexual predators. All our experiments were performed on the PANC dataset (Vogt et al. 2021).

Data. A main challenge in addressing the sexual predators' identification task through ML comes from the lack of publicly available labelled and realistic datasets. Only two distinct datasets are available: the PAN12 dataset (Inches and Crestani 2012) and the ChatCoder2 dataset (Mcghee et al. 2011), and both take their grooming examples from the Perverted Justice (PJ) website[2]. The PAN12 dataset was created by sampling non-grooming examples from IRC logs and the Omegle forum[3] and grooming examples from the PJ website. Every conversation was then split into segments of 150 messages or less assigned to a user. The ChatCoder2 dataset, on the other hand, only contains grooming examples: 497 complete conversations also extracted from PJ. Since the eSPD task requires both complete conversations as well as grooming and non-grooming examples, Vogt et al. (2021) introduced the PANC dataset as a better alternative for the task. They merged the non-grooming examples from the PAN12 dataset with the grooming examples from ChatCoder2. To standardize the dataset, they also split every grooming conversation into multiple segments of 150 messages or less assigned to the same user. Table 4 in Appendix A presents statistics about the dataset.

Inference of an eSPD System. The considered inference phase is similar to Vogt et al. (2021)'s framework for eSPD. We use a sliding window for the 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 $t_s, t_{s+1}, \ldots, t_{l-1}$, at step $s+1$ the classifier labels the sequence $t_{s+1}, t_{s+2}, \ldots, t_l$.

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. An eSPD system never classifies a conversation as non-grooming if there are messages left, or if it is still ongoing. Figure 3 in Appendix B illustrates this framework.

Evaluation Metrics. We use the latency-weighted F1 score (Sadeque et al. 2018; Vogt et al. 2021) to evaluate our models. The F-latency score measures the trade-off between the speed of detection (i.e. how early in a conversation grooming is detected) and the accuracy of the warning by applying a penalty that increases with the warning latency.

---

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

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

The warning latency is defined as the number of messages exchanged before a warning is raised (Vogt et al. 2021). As such, *a higher F-latency score means a better-performing eSPD system.* 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.

The *speed* of an eSPD system over a test set of grooming conversations is defined as speed $= 1 - \text{median}\{\text{penalty}(l) \mid l \in \text{latencies}\}$ where *latencies* corresponds to the list of warning latencies produced by the system for all grooming conversations for which a warning is raised. We can then formally define F-latency as: $\text{F-latency} = \text{F1} \cdot \text{speed}$.

While F1 is computed across the entire test set of positive and negative messages, penalty and speed are computed for the positive conversations only. This is common practice in the literature as 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).

Choice of the classifier. Although fine-tuning BERT (Devlin et al. 2019) has been shown to give better results for eSPD (Vogt et al. 2021) (see Appendix E), we use the pre-trained feature-based approach with logistic regression (LR) for our experiments since it is far less computationally expensive and better suited for scaling federated training to a large number of clients. 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 LR with a binary cross entropy loss function. For each user's segment, we obtain a 768 length vector.

6 Empirical Results and Discussion

Details about our implementation can be found in Appendix C. In this section, we investigate three research questions.

6.1 RQ1: How is the Utility of the eSPD System Affected by the FL framework?

We compare the utility of our *FL model* with two baselines: (1) *Warm-up model*: A BERT+LR model trained on the warm-up data only, to ensure that the federated frameworks are not too biased by the data-sharing strategy; and (2) *Centralized model*: A BERT+LR model trained centrally on the training data and the warm-up data.

In Table 1, we see that the federated model shows competitive results for the eSPD task with the highest F1 score (82%) and F-latency score (64%). The FL model also have the lowest false positive rate (FPR) with 3% showing that the FL framework is indeed suited for the accurate detection of predators. Finally, despite having a very high speed, the baseline *Warm-up model* has the lowest F-latency score: our model is therefore not biased by the data-sharing strategy and it is indeed learning from each client's personal data. Additional qualitative results are presented in Appendix D.

6.2 RQ2: How to Reduce the Harm of False Positives in eSPD?

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 score depends on both the F1-score and the speed. And while the F1-score takes into consideration both recall and precision, detection speed comes at the cost of precision as shown in Table 1. In a real-life setting, we expect an alarm to be raised each time a predator is detected. Given the nature of grooming messages, it is more likely that an eSPD system will be mistaken when confronted with sexual or intimate conversations, which could lead to innocent people being wrongly accused (e.g. sex workers).

Furthermore, in a real-life scenario, data often originates from forensic evidence after a predator's conviction. Given the over-representation of certain demographic groups in prison because of the failures of the criminal justice system (Government of Canada 2023a,b), we could imagine a model making spurious correlations between vernacular and predatory behavior, leading to biases against specific demographic groups. Therefore, we propose an approach to consider the cost of falsely accusing someone as a predator. For this purpose, for each of our models, we identify the classification threshold needed to achieve a 1% FPR when evaluated on the test set. Using this new threshold, we re-evaluate our models. Table 1 shows that varying the threshold comes with a loss in speed, which is to be expected since higher prediction scores are now needed to classify a window as OG. Furthermore, the results for the baseline *Warm-up model* are not presented because the smaller FPR attained for this model with a 0.99 classification threshold is 9%: the model is therefore falsely classifying non-OG conversations as OG. Finally, we notice a decrease in F-latency for all the models, a necessary trade-off to achieve better precision.

6.3 RQ3: What are the Benefits and Shortcomings of Different Privacy Implementations for eSPD?

Language data, especially personal conversations on social media, is very sensitive for two main reasons. First, it contains a lot of personally identifiable information (PII): people share their name, number, address, and personal details about themselves through text messaging. Second, studies have shown that beyond PII, it was possible to recover information about the author of a text based on linguistic patterns (Mattern, Weggenmann, and Kerschbaum 2022). For these reasons, and given the sensitivity of the eSPD application, we look at how different privacy implementations impact our eSPD system and what it means in a real-life scenario. In this work, we have discussed four different ways to offer privacy protections in the context of eSPD, and each comes with its benefits and shortcomings.

The FL implementation. Even though this implementation does not guarantee privacy, the federated framework of-

Model	F1	Recall	Precision	Speed	F-latency	FPR
Warm-up model	0.50	**0.98**	0.33	**0.96**	0.48	0.24
Centralized model	0.75	0.95	0.62	0.83	0.63	0.07
FL model	**0.82**	0.85	**0.79**	0.79	**0.64**	**0.03**
Evaluation results for a 1% FPR						
Warm-up model	-	-	-	-	-	-
Centralized model	**0.85**	**0.83**	0.88	0.69	0.59	0.01
FL model	0.83	0.78	**0.89**	**0.73**	**0.61**	0.01

Table 1: Evaluation results for the eSPD task

	Privacy Protection				**Utility**					
Model	**Input**	**Output**	**Type**	**Level**	**F1**	**Recall**	**Precision**	**Speed**	**F-latency**	**FPR**
FL with metric DP	Input	Output	Local	Instance	0.73	0.64	0.84	0.71	0.52	0.02
FL with DP-SGD	Input	Output	Local	Instance	0.77	0.86	0.70	0.79	0.61	0.04
FL with DP-FedAvg	Input	Output	Global	User	0.81	0.86	0.76	0.78	0.63	0.03

Table 2: Evaluation results for the eSPD task with different privacy implementations

fers additional protection compared to the normal centralized setting since each user's raw data never leaves their device. However, information about the training data can still be recovered from the trained model. Given the sensitive nature of personal conversations on social media, a federated framework alone may not be enough for eSPD.

The FL with metric DP implementation. We first consider a scenario where each user adds noise to their private data. As such, even during local training, the service provider never has access to the raw data. Table 2 shows that the *FL model with metric DP* has a 52% F-latency with an $\eta = 20$. Even with our oversampling technique during training, injecting noise directly into the training data comes with a high utility cost. In Figure 6 in Appendix C, we can see that models trained with more restrictive privacy budget (e.g. $\eta = 10$) are non-performing (e.g. with an F-latency score of 6%). This implementation however offers strong privacy guarantees since it gives no access to the raw data even for local training. However, metric DP might not be the most appropriate implementation in the case of eSPD. Indeed, bounding the privacy budget to a distance measure still gives information about the neighborhood of the original input: a sexual message for example is still a sexual message if not enough noise is added to the data. Simultaneously, having a more restrictive privacy budget have a high cost on the utility of the eSPD system since it means increasing the semantic distance between the original input and the perturbed one.

The FL with DP-SGD implementation. To add formal privacy guarantees to the FL framework, we implement DP-SGD. DP-SGD offers strong privacy protection against membership inference attacks by perturbing the training of a model. While in this scenario, the service provider still has access to the raw data, the data never leaves a user's device, and the updates aggregated for federated training are private. By training privately on each user's device, we guarantee instance-level DP. In our case, the instances being each data point shared by the users. This implementation also comes with a cost in utility. We experiment with different epsilon bounds at the users' levels and we see that with ($\epsilon = 0.50, \delta = 10^-5$) we register a drop in utility with a 57% F-latency score. In Table 2, we can see that with ($\epsilon = 1, \delta = 10^-5$) we can still achieve a f-latency score of 61%. It is important to note however that the privacy budget computed is not the total privacy budget for the implementation, since the same client can be resampled. In text classification with FL, each user is expected to share thousands of data points, and instance-level DP only offers plausible deniability that each data point was not part of the training rather than protecting the user in itself. This might not be the most appropriate implementation for eSPD, since we expect a child's entire conversation history to be protected, instead of individual messages.

The FL with DP-FedAvg Implementation. In DP-FedAvg, each user train a model locally and send non-private updates to the server. The server aggregates these updates and makes them private by adding noise. This implementation rely on the assumption than the server is trustworthy and that the updates are not intercepted. Training with DP-FedAvg usually has less of an impact on the utility of the model. Indeed, by adding noise after the training is done, the DP-FedAvg algorithm ensures privacy guarantees with almost no utility loss as we can see in Table 2 with an F-latency score of 62% for a total privacy budget of ($\epsilon = 1, \delta = 10^-5$). Indeed, it has been shown that given sufficient local computation on each client's data, DP-FedAvg has a negligible impact on utility (McMahan et al. 2017b), as such, we can achieve similar utility by sampling fewer clients at each round. Finally, training with DP-FedAvg ensures user-level protection since it offers privacy guarantees for all the data points shared by one user, which is essential in the context of eSPD. Indeed, since each user may contribute thousands of data points, having instance-level protections may not be enough. However, this approach does not protect the clients from the server since non-private updates are exchanged. Furthermore, as DP guarantees rely on amplification-via-sampling, a large number of clients is needed for a DP-FedAvg model to achieve good utility, which may be problematic in our case since labeled data is scarce in eSPD.

7 Conclusion and Future Directions

In December 2023, Meta announced the implementation of end-to-end encryption for one-to-one messages and voice calls on Messenger and Facebook (Goggin 2023), a response to public backlash following its cooperation in a Nebraska abortion case where personal chat data was turned over to the police (Collier and Burke 2022). This decision has faced opposition from child safety groups, who view end-to-end encryption as a setback to detecting online grooming on social media (Goggin 2023). However, EU lawmakers do not consider end-to-end encryption and child protection as conflicting, evident in their regulations compelling major tech companies to identify and remove online child abuse material (Chee 2023). Finding alternatives to current privacy-invasive monitoring systems is therefore critical to ensure children's safety. In this work, we propose different federated learning frameworks for the early detection of sexual predators as a solution for both the lack of available labelled datasets and the privacy challenges that come with such an application. Indeed, by combining different implementations of differential privacy in our framework, we show that children's safety does not necessarily mean sacrificing privacy. We aim for our work to highlight the significance of this application and encourage the creation of more datasets and research in this area. Future research could also examine the integration of various privacy-enhancing technologies to secure against the possibility of encryption backdoors or client-side scanning. This can be achieved through techniques such as secure aggregation (Li et al. 2021) or homomorphic encryption (Wibawa et al. 2022). Furthermore, we believe that our framework has the potential for broader application, such as in the early detection of cyberbullying or depression.

Limitations and Ethical Considerations

One of the main challenges of the eSPD task comes from the lack of publicly available labeled and realistic datasets. The different datasets used in the literature all 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, Chiang, and Grant 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, as opposed to conversations among family members, friends or partners. We hope that the frameworks we propose in this paper will give access to a larger range of training examples.

Involving law enforcement could also have disastrous consequences. The resulting model could be biased towards certain populations like sex workers, people from the LGBTQI+ community, or people prone to online dating. Furthermore, language models have been shown to reproduce racial and gender biases (Liang et al. 2021). As such, using such models as a basis for identifying potential suspects to be prosecuted could lead to unanticipated outcomes. Such a system should therefore never be used directly by law enforcement agencies at the risk of exacerbating existing social inequalities and persecuting innocents.

Acknowledgments

This project was partially funded by Mila AI for Humanity, Canada CIFAR AI chair, FRQNT grants and UW Global Innovation Fund. We also express our gratitude to Compute Canada and Mila clusters for their support in providing facilities for our evaluations. We would also like to thank Dr. April Edwards for sharing with us the ChatCoder 2 dataset and Mathias Vogt for taking the time to answer our questions regarding their early detection framework.

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.

Adams, S.; Melanson, D.; and De Cock, M. 2021. Private Text Classification with Convolutional Neural Networks. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, 53–58.

Agarwal, N.; Ünlü, T.; Wani, M. A.; and Bours, P. 2021. Predatory Conversation Detection Using Transfer Learning Approach. In *Machine Learning, Optimization, and Data Science: 7th International Conference, LOD 2021, Grasmere, UK, October 4–8, 2021, Revised Selected Papers, Part I*, 488–499. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-95466-6.

Basu, P.; Roy, T. S.; Naidu, R.; and Muftuoglu, Z. 2021. Privacy enabled Financial Text Classification using Differential Privacy and Federated Learning. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, 50–55. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Brown, H.; Lee, K.; Mireshghallah, F.; Shokri, R.; and Tramer, F. 2022. What Does it Mean for a Language Model to Preserve Privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2280–2292.

Canada, P. S. 2024. About Online Child Sexual Exploitation.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2022. Quantifying Memorization Across Neural Language Models.

Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; and Song, D. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, 267–284.

Chee, F. Y. 2023. EU lawmakers back rules forcing Big Tech to tackle child pornography. https://www.reuters.com/technology/eu-lawmakers-back-rules-forcing-big-tech-tackle-child-pornography-2023-11-22/. Accessed: January 8, 2024.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Dwork, C. 2008. Differential Privacy: A Survey of Results. In Agrawal, M.; Du, D.; Duan, Z.; and Li, A., eds., *Theory and Applications of Models of Computation*, 1–19. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-79228-4.

Escalante, H. J.; Montes y Gomez, M.; Villasenor, L.; and Errecalde, M. L. 2016. Early text classification: a Naïve solution. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 91–99. San Diego, California: Association for Computational Linguistics.

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

European Parliament. 2024. Child sexual abuse online: current rules extended until April 2026.

Feyisetan, O.; Balle, B.; Drake, T.; and Diethe, T. 2020. Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, 178–186. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368223.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333.

Goggin, B. 2023. Meta defaults Facebook and Messenger to end-to-end encryption despite child safety concerns.

Government of Canada. 2023a. Overrepresentation of Black People in the Canadian Criminal Justice System.

Government of Canada. 2023b. Overrepresentation of Indigenous People in the Canadian Criminal Justice System: Causes and Responses.

Hill, K. 2012. Yes, Facebook Scans Users' Private Conversations Looking For Sexual Predators and Child Porn.

Hill, K. 2022. A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal.

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.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K. A.; Charles, Z.; Cormode, G.; Cummings, R.; D'Oliveira, R. G. L.; Rouayheb, S. E.; Evans, D.; Gardner, J.; Garrett, Z.; Gascón, A.; Ghazi, B.; Gibbons, P. B.; Gruteser, M.; Harchaoui, Z.; He, C.; He, L.; Huo, Z.; Hutchinson, B.; Hsu, J.; Jaggi, M.; Javidi, T.; Joshi, G.; Khodak, M.; Konečný, J.; Korolova, A.; Koushanfar, F.; Koyejo, S.; Lepoint, T.; Liu, Y.; Mittal, P.; Mohri, M.; Nock, R.; Özgür, A.; Pagh, R.; Raykova, M.; Qi, H.; Ramage, D.; Raskar, R.; Song, D.; Song, W.; Stich, S. U.; Sun, Z.; Suresh, A. T.; Tramèr, F.; Vepakomma, P.; Wang, J.; Xiong, L.; Xu, Z.; Yang, Q.; Yu, F. X.; Yu, H.; and Zhao, S. 2019. Advances and Open Problems in Federated Learning. *CoRR*, abs/1912.04977.

Klymenko, O.; Meisenbacher, S.; and Matthes, F. 2022. Differential Privacy in Natural Language Processing The Story So Far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*.

Lee, G.; et al. 2022. Privacy-Preserving Text Classification on BERT Embeddings with Homomorphic Encryption. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3169–3175. Seattle, United States: Association for Computational Linguistics.

Li, K. H.; de Gusmão, P. P. B.; Beutel, D. J.; and Lane, N. D. 2021. Secure Aggregation for Federated Learning in Flower. In *Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning*, DistributedML '21, 8–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391344.

Liang, P. P.; et al. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, 6565–6576. PMLR.

Liu, D.; et al. 2017. A Novel Way of Identifying Cyber Predators. *CoRR*, abs/1712.03903.

López-Monroy, A. P.; González, F. A.; Montes, M.; Escalante, H. J.; and Solorio, T. 2018. Early Text Classification Using Multi-Resolution Concept Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1216–1225. New Orleans, Louisiana: Association for Computational Linguistics.

Lorenzo-Dus, N.; et al. 2016. Understanding grooming discourse in computer-mediated environments. *Discourse, Context & Media*, 12: 40–50.

Losada, D. E.; et al. 2020. Overview of ERisk 2020: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, 272–287. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58218-0.

Mattern, J.; Weggenmann, B.; and Kerschbaum, F. 2022. The Limits of Word Level Differential Privacy. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 867–881. Seattle, United States: Association for Computational Linguistics.

Mcghee, I.; Bayzick, J.; Edwards, A.; Edwards, L.; Mcbride, A.; and Jakubowski, E. 2011. Learning to Identify Internet

Sexual Predation. *International Journal of Electronic Commerce*, 15.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017a. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282. PMLR.

McMahan, H. B.; Ramage, D.; Talwar, K.; and Zhang, L. 2017b. Learning Differentially Private Language Models Without Losing Accuracy. *CoRR*, abs/1710.06963.

Morgan, J.; Paiement, A.; Lorenzo-Dus, N.; Kinzel, A.; and Cristofaro, M. D. 2021. Integrating linguistic knowledge into DNNs: Application to online grooming detection.

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

Popescu, M.; and Grozea, C. 2012. Kernel Methods and String Kernels for Authorship Analysis. In *CLEF (Online Working Notes/Labs/Workshop)*. Citeseer.

Qu, C.; Kong, W.; Yang, L.; Zhang, M.; Bendersky, M.; and Najork, M. 2021. Natural Language Understanding with Privacy-Preserving BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 1488–1497. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.

Radauskas, Gintaras. 2023. EU's plans to combat online child abuse pose risk to encryption – interview.

Reich, D.; Todoki, A.; Dowsley, R.; De Cock, M.; and Nascimento, A. 2019. Privacy-Preserving Classification of Personal Text Messages with Secure Multi-Party Computation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 3752–3764.

Resende, A.; Railsback, D.; Dowsley, R.; Nascimento, A. C. A.; and Aranha, D. F. 2022. Fast Privacy-Preserving Text Classification Based on Secure Multiparty Computation. *IEEE Transactions on Information Forensics and Security*, 17: 428–442.

Rezaee Borj, P.; Raja, K.; and Bours, P. 2023. Detecting Online Grooming By Simple Contrastive Chat Embeddings. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*, IWSPA '23, 57–65. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700996.

Sadeque, F.; et al. 2018. Measuring the Latency of Depression Detection in Social Media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, 495–503. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355810.

Samee, N. A.; Khan, U.; Khan, S.; Jamjoom, M. M.; Sharif, M.; and Kim, D. H. 2023. Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection. *IEEE Access*.

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.

Shetty, N. P.; Muniyal, B.; Priyanshu, A.; and Das, V. R. 2023. FedBully: A Cross-Device Federated Approach for Privacy Enabled Cyber Bullying Detection using Sentence Encoders. *Journal of Cyber Security and Mobility*, 12(04): 465–496.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18.

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

Song, C.; Ristenpart, T.; and Shmatikov, V. 2017. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 587–601.

Thakkar, O. D.; et al. 2021. Understanding unintended memorization in language models under federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, 1–10.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, 601–618.

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)*.

Vincent, J. 2022. New EU rules would require chat apps to scan private messages for child abuse. *The Verge*.

Vogt, M.; et al. 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)*, 4985–4999. Online: Association for Computational Linguistics.

Wibawa, F.; Catak, F. O.; Sarp, S.; Kuzlu, M.; and Cali, U. 2022. Homomorphic Encryption and Federated Learning based Privacy-Preserving CNN Training: COVID-19 Detection Use-Case.

Zambrano, P.; et al. 2019. Technical Mapping of the Grooming Anatomy Using Machine Learning Paradigms: An Information Security Approach. *IEEE Access*, 7: 142129–142146.

Zhao, Y.; et al. 2018. Federated Learning with Non-IID Data. *CoRR*, abs/1806.00582.

Zhu, H.; Xu, J.; Liu, S.; and Jin, Y. 2021. Federated Learning on Non-IID Data: A Survey. *CoRR*, abs/2106.06843.

Zhu, X.; Wang, J.; Hong, Z.; and Xiao, J. 2020. Empirical studies of institutional federated learning for natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 625–634.