SSC Annual Meeting, May 2022
Proceedings of the Survey Methods Section

Privacy Enhancing Technologies at Statistics Canada

Christopher Dugdale[1] Saeid Molladavoudi[2] Benjamin Santos[3] Julian Templeton[4]

ABSTRACT

Privacy Enhancing Technologies (PETs) are an emerging class of technologies with a promise to protect the privacy and confidentiality of data throughout its life cycle, while maintaining its utility. PETs provide Statistical Offices opportunities to facilitate collaborative analytic on less-accessible data to derive valuable insights. Statistics Canada has started experimenting with PETs a few years ago. To this end, multiple research projects have successfully been completed, such as the application of homomorphic encryption on training a machine learning (ML) classifier, privacy preserving record linkage with secure Multi-Party Computation and applying Federated Learning in the context of privacy preserving crowdsourcing. In this article, we will discuss some of these activities and share insights on potential opportunities and challenges of adopting PETs in the Official Statistics.

KEY WORDS: Privacy enhancing technologies, homomorphic encryption, secure multi-party computation, federated learning, official statistics

RÉSUMÉ

Les technologies d'amélioration de la confidentialité (TAC) représentent une nouvelle catégorie de technologies prometteuses pour la protection de la vie privée et la confidentialité des données tout au long du cycle de vie de ces dernières, tout en conservant leur utilité. Les technologies d'amélioration de la confidentialité permettent aux bureaux de statistique de faciliter l'analyse collaborative de données moins accessibles afin d'en tirer de précieux renseignements. Statistique Canada a commencé à expérimenter les TAC il y a quelques années. À cette fin, de nombreux projets de recherche ont été menés à bien, tels que l'application du chiffrement homomorphe à la formation d'un classificateur d'apprentissage automatique (AA), le couplage d'enregistrements préservant la confidentialité avec le calcul multipartite sécurisé et l'application de l'apprentissage fédéré dans le contexte de la production participative préservant la confidentialité. Dans cet article, nous présenterons certaines de ces activités et nous échangerons nos points de vue sur les possibilités et les défis potentiels de l'adoption des TAC dans les statistiques officielles.

MOTS CLÉS : Technologies d'amélioration de la confidentialité, chiffrement homomorphe, calcul multipartite sécurisé, apprentissage fédéré, statistiques officielles.

1 INTRODUCTION

In today's world, data flows in every direction in such a way that signals and noise are often indistinguishable. The data volumes are going to grow in the coming years and decades with emerging technological advancements. It is not just the volume, but also the *velocity*, *variety* and *veracity* of the generated data are orders of magnitude higher than any conceivable data collection in the pre-digital era. Examples of such data include sensors, mobile applications, satellite imagery, Internet-of-Things, 5G networks, etc. The new waves of data can create many opportunities for society and industries to leverage in the coming years; however, cybersecurity and data privacy are still among the most pressing and imperative issues.

[1]Christopher Dugdale, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6, christopher.dugdale@statcan.gc.ca.

[2]Saeid Molladavoudi, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6, saeid.molladavoudi@statcan.gc.ca.

[3]Benjamin Santos, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6, benjamin.santos@statcan.gc.ca.

[4]Julian Templeton, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6, julian.templeton@statcan.gc.ca.

To understand where the main problem lies, we have to first review the life-cycle of the data. More precisely, data lives in three states: at *rest*, in *transit* and in *use*. It is well known that data is vulnerable throughout its life-cycle and as a result, cybersecurity protocols for data protection at rest, for instance Symmetric Key Encryption, and in-transit, such as Transport Layer Security, have been standardized and implemented at large scale, e.g., in digital signatures. In recent years, Privacy Enhancing Technologies (PETs) have emerged to provide data protection while enabling data processing (Van Blarkom et al., 2003). In fact, PETs is a generic term that covers a broad range of approaches that promise to provide protection for data throughout its life-cycle, i.e. while collecting the data, processing it and disseminating the results. These approaches include homomorphic encryption, secure multi-party computation, differential privacy, distributed ML (e.g., Federated Learning), trusted execution environments and zero-knowledge proofs. We provide a high-level review for a few of these methods in the subsequent sections.

Statistics Canada has the mandate to provide statistical information and data about Canada's economy, society and environment to help Canadians better understand their country and improve public decision making for the benefit of all Canadians. The organization already has rigorous measures in place to preserve privacy and confidentiality in the modern digital era. As the agency continues to implement new technologies and innovations, its commitment to protecting privacy and security remains the highest priority. The data science team at Statistics Canada has been exploring the use of these existing and emerging PETs to continuously address the privacy preservation needs for highly sensitive information in various statistical programs. In addition to alternative storage options, PETs will allow the agency to adopt and implement remote and delegated computing on encrypted data, benefit from potential multi-party computation opportunities and derive insights from distributed and inaccessible data (Government of Canada, 2021).

It is important to note that PETs do not solve the important trade-off between security and privacy in one hand and data use in the other by themselves, but only offer risk mitigation that may be the difference between a statistical project being a go or no-go. Another important aspect of the PETs is that some of them are designed and proposed to address the *input privacy*, while others are focused more on the issue of *output privacy*. Input privacy is concerned with how to ensure privacy of the input data of one or more participating parties (or data holders), who enter a joint function, e.g., a statistical algorithm. Output privacy on the other hand typically relies on either aggregation or sensitivity analysis, e.g., traditional data disclosure controls, or perturbation, e.g., differential privacy. The aim here is to prevent and reduce risks of re-identification of information about data subjects, e.g., personal identifiable information, by reverse engineering the outputs of the statistical algorithms and published data. It is worth noting that Statistics Canada operates under the *Statistics Act*, as its legislative framework. As a result, for decades, the agency has effectively developed statistical disclosure control measures to satisfy the requirements of the Statistics Act, ensuring that no sensitive personal or micro-data is disclosed while disseminating statistical products.

In this article, we will review a few of the research projects at Statistics Canada that involve PETs, with a focus on those that address the input privacy. In section 2, we will cover one application of homomorphic encryption on a delegated computing scenario, where data providers, data consumers and the computing party are all distinct entities. Next, we will provide an overview of a two-party privacy preserving record linkage protocol in section 3, where a combination of data obfuscation and secure multi-party computation is used. Then, we will review an application of Federated Learning (FL) approach in the context of a privacy preserving crowdsourcing activity in section 4. Finally, in section 5 we will provide conclusions and outlooks of the direction that the data science group in Statistics Canada is taking in this area. It is worth noting that none of these projects are in production at Statistics Canada at the time of writing this article.

2 Supervised text classification with leveled homomorphic encryption

At a high-level, Homomorphic Encryption (HE) is an asymmetric crypto-system that allows arithmetic operations to be performed on encrypted data without the need to decrypt it first. The core idea behind HE is to make the encryption map a ring homomorphism from the plaintext space to the ciphertext space, which would then preserve addition and multiplication operations. This is unlike the existing standard encryption protocols that would first require decrypting the ciphertexts with the private key. In the last few years, focused research on HE has resulted in reducing its computational and communication costs by orders of magnitudes, while providing cryptographic security. On the application front, HE has been applied to delegated computing scenarios involving sensitive data in areas such as health, finance, and justice (Raisaro et al., 2018).

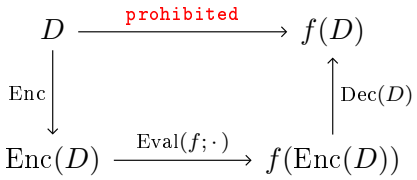

Figure 1: High-level graphical representation of a HE-based cloud computation scenario. We have a private data set D that we would like to apply a function f to, but for some reasons this is prohibited. Perhaps f is computationally expensive, or proprietary. With HE, we can encrypt our data to $\mathrm{Enc}(D)$ and send it to the cloud, who can apply f homomorphically, i.e. $f(\mathrm{Enc}(D))$, and then return it to us to decrypt and use. We get our desired results at the end of the protocol, i.e. $f(D)$, without the cloud having access to the data D during or after it.

In the first research project on PETs, we consider an HE cloud computing paradigm where data providers and consumers are separate entities (see Figure 1). To be precise, in this hypothetical paradigm the data providers would be retailers across Canada, who own the *scanner data*, encoding point-of-sale transactions and consisting of product description, some identifiers and prices. Scanner data, which is statistically sensitive information, is currently used in business statistics programs at Statistics Canada to produce various price statistics. HE would allow Statistics Canada to outsource part of the scanner data workflow to the cloud, while ensuring that the privacy of the input data is preserved. In this scenario, Statistics Canada would be the consumer party that would delegate part of its internal workflow to the cloud, as the computing party, while preserving the privacy requirements of the input data (see Figure 2).

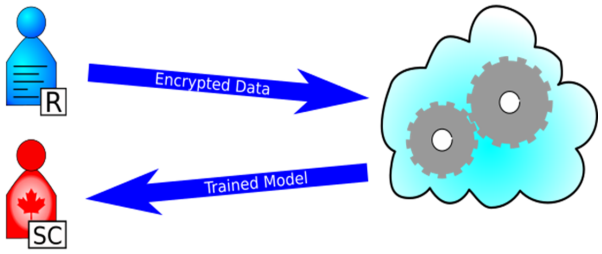

Figure 2: The HE-based cloud computation scenario, with the retailers (R) as data providers and Statistics Canada (SC) as the consumer. The cloud, as the computing party will perform the instructed operations on encrypted data, e.g., training a ML model.

In this research project we have accomplished two main tasks. The first involves the computation of simple statistics, such as the total, mean and variance of thousands of encrypted lists of synthetic price values by using a leveled HE scheme (Fan & Vercauteren, 2012). Qualitative results show roughly $\sim$ 10x run time expansion in the ciphertext space comparing to the cleartext. The second task involves the training of a single-layer neural network classifier on encrypted text data to predict the North American Product Classification System (NAPCS). The classification of the product description is in fact the first step in the scanner data's workflow. NAPCS is an international standardized system of product codes that is used to classify different types of products for the purpose of producing aggregate product statistics. In this project, we consider a similar data set containing about $50,000$ text entries from the USDA's FoodData Central (USDA, 2020) that has been manually labeled according to 5 different NAPCS codes.

Our main goal was to investigate the feasibility of using HE in computationally intensive ML tasks, such as training a neural network while preserving the confidentiality of the input data set. With techniques such as packing and multi-threading we managed to train an ensemble neural network that learns from a large encrypted data set for the supervised text classification. Comparing to the cleartext experiments, our results of experiments in the ciphertext domain prove that the performance degradation introduced by the inherent noise as well as the approximate computation of HE is manageable. HE offers an unparalleled level of cryptographic security, but, of course, adds the cost of higher computational and storage requirements. Also, the leveled nature of these schemes limits the number of consecutive operations we can perform on a ciphertext. This is because noise accumulates in a ciphertext as you compute on it until the noise finally overcomes the signal. In practice, we have about

30 consecutive multiplications on a ciphertext before we need to decrypt it. Further details about this project, including the types of models and their performance can be found in Zanussi et al. (2021).

3 Privacy preserving record linkage

Our second PET project involves record linkage, which is the process of finding records of the same units, e.g., the intersection, in two privately owned tabular data sets held by separate hypothetical entities, while performing analysis on it. Normally, this process requires at least one party who must share their private data with the other for the purpose of record linkage. However, privacy concerns and regulations often prohibit data holders to enrich their data with other sources through record linkage. In this project, we use a combination of data obfuscation methods and Secure Multi-Party Computation (SMPC) to investigate the feasibility of a privacy preserving record linkage (PPRL), namely finding the set intersection that is enriched with auxiliary information (i.e. *payload*) from the other data set and calculating some basic aggregates on it, while preserving the privacy of the input data. PPRL can have applications that are relevant to a National Statistical Office (NSO); for instance, it can facilitate the dissemination of sensitive information by reducing the administrative overhead in data sharing processes and fostering collaboration among data holders in either the same or under different jurisdictions. For a general overview of record linkage and its technical aspects, interested readers can consult (Haron et al., 2016).

In this work, we consider a scenario in which a secondary entity would like to enrich their data by performing a record linkage with survey data that is already collected by an NSO and further compute some basic aggregates on the intersection, all in a privacy preserving manner, in the sense that the external party would not have access to the survey micro-data. Here, we implement a deterministic record linkage (exact matching) on a common identifier between the two data sets and compute the aggregates on the complemented intersection by some of the numerical and categorical variables in the sample survey, such as age, sex at birth, marital status, etc. The result is a table of weighted sums or averages, one for each value of every attribute in the complemented linked table.

We apply the PPRL protocol outlined in Chandran et al. (2022) and Pinkas et al. (2019) to the problem outlined above. While these works consider very simple computations on the intersection (e.g., computing the cardinality, and only returning it if it is greater than an agreed upon threshold), we extend their methods to accept payloads as well as perform generic computations on these payloads. To perform the PPRL, we have developed a two-step process. First, the data sets are input into an *Oblivious Programmable Pseudo-Random Function* (OPPRF), which allows for data to be securely obfuscated in a controlled way that facilitates linkage. Next, the parties make use of SMPC to compute the aggregates. Here, the obfuscated values are split into secret shares and used to compute a suite of aggregates based on the attributes present in the NSO's survey data set. We consider both numerical attributes, where the value represents some quantity, and categorical attributes, which assigns the identifier to one of a limited number of discrete classes. The goal here is to combine the micro-data from the two sources and use the linked data set to compute the desired aggregates.

We model both organizations as *semi-honest* input parties. This means that they will follow the protocol as outlined, but will always try to infer information about the other party's data. Not only are the identifiers considered sensitive, but so is the attached micro-data. The synthetic data sets used in this experiment are sized and structured based on a typical data linkage with an NSO. In particular, the dataset held by the NSO is designed to simulate a typical survey conducted by Statistics Canada, consisting of a few categorical attributes and a sample weighting factor, which is used to estimate population-level statistics from the sample. Each record has been given a random identifier, designed to mimic numerical identifiers given to citizens or businesses in some countries. The secondary organization's dataset is simpler in design, with only a single numerical value, in addition to the identifiers. The sample survey conducted by the NSO had 60,000 respondents, while the secondary organization has a dataset of about 380,000 individuals (Zanussi & Dugdale, 2022).

3.1 Protocol

The first step in the generic protocol, after having agreed on the security parameters such as bit-length, identifiers, payload, etc., is to use *Cuckoo Hashing* to sort sets into hash tables, essentially an indexed array, where a point is put into the index corresponding to its value after hashing. In our scenario, we model the external organization as the sender and the NSO as receiver. The next step is the use of OPPRF, which allows the parties to obfuscate their data sets in a controlled but private way. The inputs to the OPPRF are the two hashed tables that the parties

have obtained after the first step and the outputs are the obfuscated identifiers and their associated attributes that are present in the intersection of the two data sets.

The final step in our protocol is a secret sharing circuit that the parties evaluate cooperatively. It is worth noting that in secret sharing, which is typically utilized in SMPC, the sensitive data is split into shares and distributed among the participating computing parties to perform joint computations on distributed private data sets. At any point, if a threshold number of parties would like to reconstruct the secret, they combine their shares of the secret to produce the result. There are several secret sharing schemes that may be used such as Shamir's polynomial scheme, which involves encoding a secret by interpolating a polynomial, Blakley's plane scheme, which involves intersecting three planes in a point, Yao's garbled circuits, and what Demmler et al. (2015) calls arithmetic and boolean schemes. For this project, we have utilized the ABY framework of Demmler et al. (2015), where the authors have found that the arithmetic format is much faster than Boolean for the performance of simple arithmetic operations, like multiplication and addition. The operations, where the share format matters the most, are multiplication, multiplexer, and comparison. We use both masking and multiplexer to mitigate the lack of conditional branching in ABY.

Table 1: The time and communication cost of the aggregation circuit. Aggregates are computed on identifiers and micro-data that have been obfuscated by the OPPRF protocols.

# Aggregates	Time (s)		Data (MB)	
	Setup	Online	Setup	Online
1	0.51	0.081	39.4	3.63
10	0.645	0.475	53.5	5.63
132	2.48	6.41	242	45.2

The results for a single aggregate, for ten aggregates, and for the full suite of 132 aggregates are given in Table 1. The circuit involves an offline setup period that can be done independently by either party once the OPPRFs have been run, and the SMPC circuit is run during an online phase where the parties need to be in direct communication to complete the second step of the protocol. We have reported the communication exchanged in these periods as well as the time taken to perform them. The protocol requires 11 communication rounds, which is the number of times that the parties must exchange information in order to run the circuit.

4 Privacy preserving crowdsourcing

Federated Learning (FL) is a method for training an ML model using a distributed data set without sacrificing input privacy (Kairouz et al., 2021). Distinct data holders each train the model, e.g., a text classifier, on their data locally (on their own devices) and send the model updates to a central authority to be aggregated. For instance, several hospitals, each in possession of chest x-ray images, collaborate to train a model for detecting lung disease without the images ever leaving each hospital's servers. At a larger scale, this may help NSOs to explore the feasibility of using privacy-preserving crowdsourcing for application to surveys on different subject matter areas, including those where the ability to collect data from users is difficult. Similarly, users may be fine with having their data stored at one point in time but may wish to revoke access at another point in time. Thus, the question arises as to how an NSO can work with sensitive data without viewing or acquiring it.

The idea is that an NSO can hold and host a centralized ML model on a server and have client-facing applications, such as web pages or mobile apps, to allow clients to input their data to be trained on, without the data ever leaving the respondents' devices. Once respondents train the ML model locally, the central authority (i.e. an NSO) can retrieve only the model weights to then aggregate into an updated centralized model. This process is repeated such that a robust ML model can be trained and stored without needing to access or store any raw data from the respondents. For sensitive topics, this will allow ML models to be trained on the private data without viewing or accessing the data. Data pre-processing, such as cleaning and transformation, can be done prior to training on the respondents' devices.

In this research project, we explore the feasibility of applying FL scenarios in a non-probabilistic crowdsourcing survey, where the respondents' data never leaves their devices, e.g., mobile phones. The question we are focusing on is: *What if data owners can keep their data while a central authority, such as an NSO, is still able to use it?*. From a privacy perspective, a privacy-preserving crowdsourcing framework must allow the central authority to do analytical work (e.g., train ML models) without needing to fetch raw data from respondents. From a security

perspective, this framework should be robust against attacks aimed at discovering the trained model, unless access is given, and against gaining insights on the respondents' data that has been used for training the model. This is where the value of using privacy-preserving techniques that enforces trust, security, and privacy arises.

We develop a FL simulated environment with a publicly available data set on a social topic, i.e. cyberbullying, and conduct various tests to demonstrate its effective operation within various crowdsourcing approaches. Two frameworks are explored to evaluate how FL can be used on cyberbullying data within a crowdsourcing setting. The first, an annotator framework, which indicates that using FL with a set of trusted data annotators can allow an NSO to train a model with the annotated data while keeping the data on the annotator's side. The second framework, a Semi-Supervised Federated Learning (SSFL) framework, allows unlabeled data to be used for training while remaining on client devices.

Finding a good dataset for cyberbullying classification is challenging, as Emmery et al. (2021) states, the samples are under-powered in terms of accurately representing the substantial variation between social media platforms, the positive instances are biased; only reflecting a limited dimension of cyberbullying, and crowdsourcing bullying content potentially decreases the influence of domain-specific language-use. In Van Hee et al. (2015, 2018), the authors compiled posts from `ask.fm` using web-scrapping, where users interact with each other by posting questions and receiving answers. The researchers started with a seed profile list, which was used to select related user profiles and fetched all the posts for the augmented list. The corpus was then independently annotated by trained linguists, where specific guidelines were designed to provide annotation rules to classify cyberbullying activity, severity, category and roles. In our project, we use this data made available in an open repository that only contains the cyberbullying flag (yes or no). The encoded data was already pre-processed from raw text to numerical features, namely tokenisation and stemming were already done, resulting in 871,044 features of the following types: Word and character n-gram bag-of-words, term lists, subjectivity lexicon features and topic model features. This dataset is heavily imbalanced, with less than 5% from 113,694 data points (posts) flagged as cyberbullying, which is also a common problem of cyberbullying data sets.

4.1 Annotator Framework

This project focuses on employing FL in a crowdsourcing survey on a social topic to protect the privacy of the input data, while allowing the study of the phenomenon. FL is a class of protocols which aims to train a ML model, in particular a neural network, on input data that is owned by multiple parties (e.g., respondents), who want to keep their data private. In FL, there is a central, not-fully-trusted, authority or server (e.g., crowdsourcing survey owner, such as an NSO), who will assist the decentralized parties to train the ML model. More precisely, in FL, each party holds an identical local copy of the neural network to train. They each perform one round of training on their local devices, consisting of one or more epochs, with their private data and only send their model updates (i.e. weights) to the authority. The authority coordinates these incoming gradients, which can be as simple as aggregating them, and possibly instructs each of the parties on how to update their local models by combining the insights gained by every party's data. The process then repeats for the desired number of rounds, until the authority (and possibly every party) holds a trained version of the network. It is worth noting that the final network and the training process reveals no more about the input data than the sequence of gradients computed by each party (e.g., the respondents). Therefore, the respondents' sensitive data will always stay with them and never be transferred to the central authority/server.

The first framework that is explored is the annotator FL framework in which trusted annotators, who are subject matter experts regarding the selected domain, such as cyberbullying, can annotate samples which they will receive/hold, without disclosing the data to the central authority, who is looking to train an ML model on that data. As an example, an NSO may wish to train an ML model from data which multiple organizations, e.g., research institutes, collect without seeing the sensitive data (for legal and privacy reasons). Thus, this framework assumes that the other organizations can annotate their data and the NSO can request local training to be performed by each organization to achieve training from each sensitive data source while maintaining both user privacy and the organization's legal obligations.

Following the annotations, a server will request for the annotated data to be trained on the annotator's devices to keep the data with the trusted annotators and to never provide the data to the remote server. The annotators will train a model starting with the up-to-date weights which are stored on the server. The remote server will receive the weights from the trained ML models following the local annotators' training, which are aggregated to

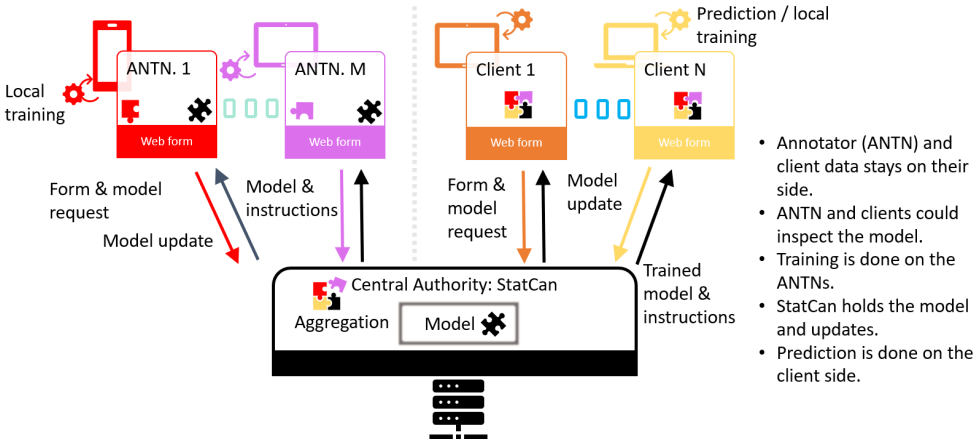

Figure 3: The annotator framework in the privacy preserving crowdsourcing with federated learning.

then be stored on the server. This way, future requests for a classification from the central model will use the aggregated weights trained from the annotators. The annotator framework, as described above, is visualized in Figure 3. Many different techniques, such as homomorphic encryption and differential privacy, can be used in combination with FL to further increase the privacy of the data being trained with (with a potential corresponding cost to the communication, run time and memory usage of the centralized model).

To compare the effectiveness of FL when applied to this annotator framework, a baseline scenario is created in which a centralized model is trained directly from the combined labeled data contained by each annotator. This centralized approach requires access to the data directly and avoids maintaining the privacy of a user's data. Through a comparative analysis, the results can provide insights into any observable pros and cons which may occur with the FL approach (in particular, how well does the FL approach learn the minority class compared to the baseline approach). The annotator framework shows strong results (shown in Table 2) on a simple model with minimal tuning. This is a strong indication that a first step in testing implementations of FL can be to collaborate with other organizations to develop a FL scenario based on some domain in which each organization can use their data to derive a centralized robust ML model. This can provide a concrete step in integrating FL into a deployed example within the government while being easier to integrate when compared to the SSFL approach.

This still requires significant planning from all parties to come to a mutual agreement regarding the annotation strategy to be taken and the requirements of the annotators themselves. If the data is already annotated, discussions must be held to determine how the data is annotated and whether it can be used within the target annotation strategy. Under the right planning based on the target domain, the annotation strategies prepared, and the data to be annotated, both the feature space and annotation strategies can coincide among organizations. A project such as this will provide a more concrete code base to work in future projects and test beyond simulation environments, which do not consider how to integrate the code into actual products.

4.2 Semi-Supervised Federated Learning

The second framework is the semi-supervised FL approach, in which all data is unlabeled and maintained on respondents' devices. In this framework, the central authority must be able to classify the data on respondents' devices prior to training without viewing the data and without annotators. To accomplish this, semi-supervised learning (SSL) is used to assign labels to the data on respondents' devices. Once the data is labeled, the FL approach can continue as normal. Unlike the annotator approach, this approach is realistic in large-scale crowdsourcing efforts in which users may provide data for training with their devices via some web page or application that allows the training to be performed exclusively on their device, with input privacy guarantees. An overview of the SSFL framework is presented in Figure 4.

Despite the considerable number of SSL training methodologies, the approach used for this framework is simple and extendable to more complex approaches. The concepts used to derive the approach are based on the SSL overview presented by Ouali et al. (2020). Within the scope of this project, first, a small subset of collected data must be used to derive a pre-trained centralized model, which will be distributed to respondents. This initial model must be trained with quality data that can generalize well when sent to clients and should be tested with an isolated

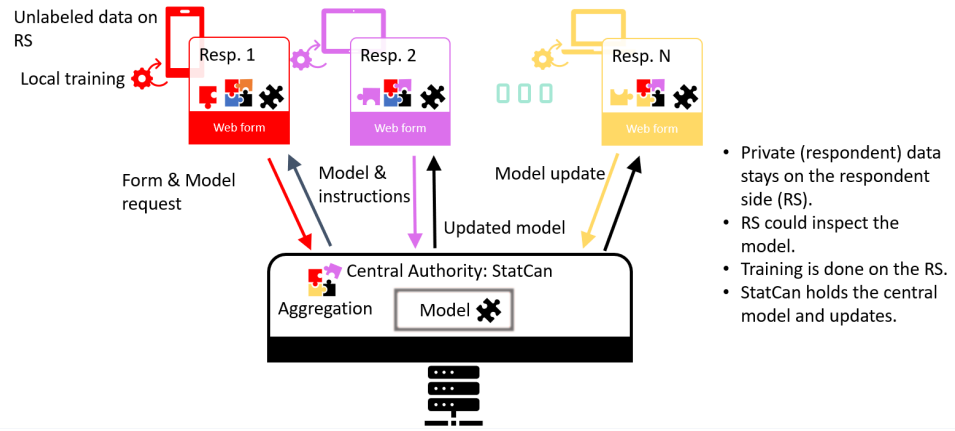

Figure 4: The Semi-Supervised Federated Learning framework in the privacy preserving crowdsourcing.

testing set, which has been collected. Alongside this initial model, an unsupervised ML algorithm should be fit with the same training data. This algorithm will be used to group samples held by respondents into clusters to help identify similarities and dissimilarities between the samples. As with the initial model, the selected clustering algorithm and the hyperparameters used for the selected model must be robust enough to create a well-trained model. The clustering algorithm will be sent to clients just as the initial model will, but will not be updated unless the central authority or trusted annotators have data that can be used to further improve the model.

With the initial models in place, the conventional FL approach is used to distribute the pre-trained model's weights to the clients for it to be further trained with their unlabeled input data on the respondents' devices. The SSL approach begins by deserializing the provided clustering algorithm and using it to cluster each sample, which can be used to train the provided model. Each clustered sample is then run through the provided ML model to receive a class prediction for the sample. With the predictions assigned to each sample and with each sample clustered, the SSL approach applies a chosen voting mechanism to assign each cluster of samples the same label. Since clusters indicate similarity between the samples, assigning one prediction to all samples in a cluster may help to remove issues when the preliminary predictions may not be completely accurate on the unseen data. This voting mechanism can be robust by using confidence outputs from the provided predictions to then use within a weighted majority vote system. However, the approach that is used within the scope of this project is a simple majority vote without using a prediction's confidence.

Now each sample will contain a label to be used for training. However, some transformations are applied to the data to avoid training the model with the initial data using the model's own predictions (which avoids learning new information). This transformed data and the corresponding labels are then used in training and the process continues for all batches and for each respondent requested to train the model. The model weights are then sent back to the server to be aggregated and the server will test how the model now performs on the isolated test set. If the change is positive it can be committed, but if the change is negative, the server can decide whether to keep the model, or store the previous model in case it is needed, or drop the newly trained model. This process is continued until a desired level of performance in achieved. This approach acts as a starting point which can be extended and tuned appropriately.

To summarize, this project explored FL as a privacy-preserving technique to allow for ML models to be trained with data that remains on the respondent's device. This approach is applied on cyberbullying data to understand its effectiveness and feasibility when applied to a task using sensitive data. Two frameworks have been explored to evaluate how FL can be used on cyberbullying data within a crowdsourcing setting, with the summarized results presented in Table 2. The first, an annotator framework, indicates that using FL with a set of trusted data annotators to allow an NSO to train with the annotated data while maintaining the data on the annotator's side has a positive impact on both performance and privacy. These strong results come with minimal tuning and a simple model, indicating stronger performance when tuned further. The second framework, an SSL FL framework, allows unlabeled data to be used for training. A basic SSL approach has been applied to the framework and is found feasible to implement. Both tuning and testing is required for this approach to be effective. A combination of the annotator and SSL approaches may yield the most success in a production environment and should be further considered. Overall, this work highlights the potential benefits of using FL, the techniques which can be used in

different applications, and the difficulties of applying the technique beyond simulation settings. In combination with other privacy-preserving techniques, FL is a significant tool that could greatly benefit NSOs to work with sensitive data, since the data does not get collected.

Table 2: Summary of results (optimal test performance) for the minority class (i.e. when the text is classified to be cyberbullying) in both the FL Annotator and Semi-Supervised Learning frameworks. In both cases, 5-fold cross validation was performed and each fold was balanced with data augmentation techniques, e.g., over or under sampling the classes.

FL Framework	Accuracy	Precision	Recall	F1-score
Annotator Framework	0.96	0.62	0.59	0.60
Semi-Supervised Learning	0.96	0.85	0.22	0.34

5 Conclusions and Outlook

In conclusion, there are many promising opportunities for PETs, as enablers, to be used at both ingestion and dissemination points at NSOs. Hence, we believe that it is an ideal time to conduct more research projects on PETs now. To unleash the power of sensitive data, PETs are being deployed in many sectors, ranging from the financial services to healthcare, pharmaceuticals, telecommunication and government bodies, etc. On the non-technical side, lack of wide-spread knowledge, for instance about their advantages, limitations, risks, costs and impact on operational routines, longer term experience with them and trust to the technology are the major roadblocks to the wider adoption of PETs. Adopting any new technology is in fact a combination of technical, legal and social aspects that will each need to be addressed on a case-by-case basis (at least in the beginning) to pave the way for large-scale adoption.

Moving forward, we plan to continue exploring PETs, e.g., multi-party PPRL and synthetic data, with the goal of moving them to production in the near future. This is a fast evolving space and according to Gartner, "by 2025, 60% of large organizations will use one or more privacy-enhancing computation techniques in analytics, business intelligence or cloud computing, an increase of 10% since last year's report" (Willemsen et al., 2022). Policies and regulations will need to evolve to envision operations in the presence of PETs, e.g., data access or data sharing, and we need to work together with those who are involved in the legal space to facilitate this vision. Last but not least, transparency, communication and social license from the public are the key factors and more work on these areas are still required to fill the gap between the legal and social licenses.

6 Acknowledgment

The authors would like to thank the sponsors, including the R&D Board at Statistics Canada and other internal and external stakeholders, for their support of these research projects. We would like to extend our gratitude to Abel Dasylva, Sevgui Erman, Zbigniew Rakowski, Eric Rancourt, Étienne Rassart, Christos Sarakinos, and last but not least Zachary Zanussi, for their valuable contributions throughout various steps of these projects.

7 Disclaimer

The content of this article represents the position of the authors and may not necessarily represent that of Statistics Canada.

REFERENCES

Chandran, N., Gupta, D. & Shah, A. (2022). *Circuit-psi with linear complexity via relaxed batch OPPRF.* Proceedings on Privacy Enhancing Technologies, **2022(1)**: 353-372.

Cheon, J.H., Kim, A., Kim, M. & Song, Y. (2017). *Homomorphic Encryption for Arithmetic of Approximate Numbers.* In Advances in Cryptology-ASIACRYPT 2017; Takagi, T., Peyrin, T., Eds. Springer International Publishing: Cham, Switzerland; pp. 409-437.

Demmler, D., Schneider, T. & Zohner, M. (2015). *ABY: A framework for efficient mixed-protocol secure two-party computation.* In Network and Distributed System Security Symposium (NDSS'15), Internet Society, San Diego, CA, USA, February 8-11, 2015.

Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V. & Daelemans, W. (2021). *Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity.* Lang Resources and Evaluation **55**, 597-633.

Fan, J. & Vercauteren, F. (2012). *Somewhat Practical Fully Homomorphic Encryption.* Cryptology ePrint Archive, Report 2012/144. Available online: https://ia.cr/2012/144 (accessed on October 1, 2022).

Government of Canada, (2021, December 15). *Data Science at Statistics Canada.* Statistics Canada. From: `https://www.statcan.gc.ca/en/data-science/stat`, (accessed on November 3, 2022).

Harron, K., Mackay, E. & Elliot, M. (2016). *An introduction to data linkage*, Admin Data Research Network.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Nitin Bhagoji, A., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B.,...Tramèr, F. (2021).*Advances and Open Problems in Federated Learning.* Foundations and Trends in Machine Learning, 14(1-2), 1–210. `https://doi.org/10.1561/2200000083)`

Ouali, Y., Hudelot, C. & Tami, M. (2020). *An overview of deep semi-supervised learning.* arXiv preprint: 2006.05278 (accessed on October 1, 2022).

Pinkas, B., Schneider, T., Tkachenko, O. & Yanai, A. (2019). *Efficient Circuit-Based PSI with Linear Communication*, pages 122-153. Advances in Cryptology – EUROCRYPT 2019. Springer, Cham, Switzerland.

Raisaro, J. L., Troncoso-Pastoriza, J. R., Misbach, M., Sousa, J. S., Pradervand, S., Missiaglia, E., Michielin, O., Ford, B. & Hubaux, J. P. (2018).*Med Co: Enabling Secure and Privacy Preserving Exploration of Distributed Clinical and Genomic Data.* IEEE-ACM transactions on computational biology and bioinformatics, **16(4)**: 1328-1341.

U.S. Department of Agriculture (USDA), A.R.S. (2020). *FoodData Central: USDA Global Branded Food Products Database.* `fdc.nal.usda.gov`.

Van Blarkom, G.W., Borking, J.J. & Olk, J.G.E. (2003). *Handbook of Privacy and Privacy-Enhancing Technologies (The Case of Intelligent Software Agents).* The Hague, College bescherming persoonsgegevens.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G.D. & Daelemans, W. (2015). *Detection and Fine-Grained Classification of Cyberbullying Events.* In International conference recent advances in natural language processing (RANLP), 672-680.

Van Hee C., Jacobs G., Emmery C., Desmet B., Lefever E., Verhoeven B., De Pauw, G., Daelemans, W. & Hoste, V. (2018). *Automatic detection of cyberbullying in social media text.* PLoS ONE **13** (10): e0203794.

Willemsen, B., Krikken, R. & Horvath, M. (2022). *Top Strategic Technology Trends for 2022: Privacy-Enhancing Computation (ID G00755920).* Retrieved from: https://www.gartner.com (accessed on October 1, 2022)

Zanussi, Z., Santos, B & Molladavoudi, S. (2021). *Supervised Text Classification with Leveled Homomorphic Encryption.* Proceedings of the 63rd ISI World Statistics Congress. https://www.isi-web.org/files/docs/papers-and-abstracts/87-day2-cps027-supervised-text-classification.pdf.

Zanussi, Z. & Dugdale, C. (2022). *Practical privacy-aware data linkage and statistical aggregation based on Privacy Enhancing Technologies*, Under preparation.