
Structured Evaluation and User Acceptance Ranking for Privacy-Preserving ML Techniques

Lucas Steinmann, Miriam Fischer

Department of Information Security,
Technical University of Munich, Munich,
Germany
{lucas.steinmann, miriam.fischer}@tum.de

Andrea Rossi, Tomoko Yamamoto

KDDI Research, Inc. Saitama, Japan
{andrea.rossi,tomoko.yamamoto}@kddi-
research.jp

ABSTRACT

Using Privacy-Enhancing Technologies (PETs) for machine learning often influences the characteristics of a machine learning approach, e.g., the needed computational power, timing of the answers or how the data can be utilized. When designing a new service, the developer faces the problem that some decisions require a trade-off. For example, the use of a PET may cause a delay in the responses or adding noise to the data to improve the users' privacy might have a negative impact on the accuracy of the machine learning approach. As of now, there is no structured way how the users' perception of a machine learning based service can contribute to the selection of Privacy Preserving Machine Learning (PPML) methods. This is especially a challenge since one cannot assume that users have a deep technical understanding of these technologies. Therefore, they can only be asked about certain attributes that they can perceive when using the service and not directly which PPML they prefer.

This study introduces a decision support framework with the aim of supporting the selection of PPML technologies based on user preferences. Based on prior work analysing User Acceptance Criteria (UAC), we translate these criteria into differentiating characteristics for various PPML techniques. As a final result, we achieve a technology ranking based on the User Acceptance Criteria while providing technology insights for the developers. We demonstrate its application using the use case of classifying privacy-relevant information.

Our contribution consists of the decision support framework which consists of a process to connect PPML technologies with UAC, a process for evaluating the characteristics that separate PPML techniques, and a ranking method to evaluate the best PPML technique for the use case.

Keywords privacy-preserving machine learning · privacy by design · privacy-enhancing technologies

1 Introduction

Which Privacy Preserving Machine Learning (PPML) technique a user would prefer to be implemented in a specific use case is a difficult question to assess from a developer's perspective. Since users lack fundamental knowledge about these technologies it is difficult to obtain preferences for implementing a PPML technique in a certain application. But for the actual use of an application by the user, meeting their needs is a crucial requirement for AI service providers. This is on the one hand driven by the need to meet legal regulations such as the GDPR. On the other hand, privacy can be used as a selling point when comparing to other competitors. Although the principles of privacy by design [18] claim full functionality in a win-win approach, in practice this is not easy to achieve for Machine Learning (ML) applications. Almost always PPML techniques such as Differential Privacy (DP), Homomorphic Encryption (HE) or Secure Multiparty Computation (SMPC), come with different trade-offs in performance, accuracy

and adaptability [45]. Moreover, these techniques are often not dominating each other but come with individual trade-offs that influence different aspects of an application. To address this research gap, we provide a structured decision support framework for developers to support them in the selection of a PPML technology that best suits the users' preferences. The result can be used on its own or in a more general PPML selection process as one decision criterion out of many. To achieve this, we built on the analysis of [45] who elicited User Acceptance Criteria (UAC) that influence the users' acceptance of an application based on frequently used models, such as IUIPC [80, 50] or APCO [7] model. We use our framework to translate these criteria into differentiating PPML Characteristics among PPML technologies.

Our contribution is the provision of a structured decision support framework for developers to rank PPML techniques based on UAC to answer the research question (RQ): "Which PPML technique would a User Choose?". To better address this RQ, we split it into three research objectives (RO).

- RO1: Providing a process on how to connect PPML technologies with UAC. We contribute by providing precise formulas to apply the suggested mapping of Löbner et al. [45] to calculate a PPML Characteristic preference score from UAC.
- RO2: Providing a process for evaluating the characteristics that separate PPML techniques. We describe how PPML Characteristics can be divided into categories. Moreover, we suggest setting weights for each category using weight vectors.
- RO3: Providing a ranking method to evaluate the best PPML technique for the use case, based on RO1 and RO2. We contribute by providing a formula that connects PPML Characteristic preference scores translated from the user input and the weighted PPML categories.

We address our research objectives by guiding the reader through all the steps of our framework. Furthermore, we showcase the use of the framework by the example of a simplified Privacy Sensitive Information (PSI) detection application. Thus, the focus is on the process not on the evaluation of the mapping itself. To be precise, based on the showcasing, we provide an explanation for the PSI detection application of how to arrive with our framework at the weighted list of PPML techniques, utilising the users' UAC as an input to the framework.

The structure of this paper is as follows: In section 2 (*related literature*) we collect papers that provide decision support to choose PPML techniques. Section 3 (*methodology*) describes relevant entities, research objectives, and the methodology used to address them. In section 4 (*framework description*) we explain the calculation steps and how to apply them, which is the major contribution of this paper. In section 5, we introduce a *use case: PSI detection in texts*, which we use as an example to demonstrate the use of the framework. In section 6 (*framework application*) we apply the steps from section 4 to the use case described in section 5. In section 7 (*discussion*) we reflect on our results and implementation considerations taking into account the limitations and suggesting avenues for future research. Finally, section 8 (*conclusion*) summarises our main findings.

2 Related Literature

There are several studies that provide classification and comparison of PPML techniques. Zheng et al. [81] classified PPML systems based on key technologies with a special focus on applying the PPML techniques to IoT. Thus, they compared the techniques according to computation and communication overhead. They first considered whether the system is for training or for inference. For privacy-preserving training, they further divided into parameter transmission based techniques (e.g., federated learning) and data transmission-based techniques (e.g., anonymization and data obfuscation). From their classification, it is difficult to see which key technologies can be combined. Tanuwidjaja et al. [66] presented a similar classification of privacy-preserving deep learning on machine learning as a service. Boulemtafes et al. [11] classified PPML systems based on the structures of the systems. They divided the system settings into four layers: (1) learning, inference, or releasing a model, (2) collaborative learning (multiple participants are involved) or individual learning (a single participant is involved), (3) server-based (most tasks are outsourced) or server-assisted (tasks are performed cooperatively between participants and the outsourcing servers), and (4) key technological concepts (e.g., encryption). Similar to [81], it is difficult to see from their classification which key technologies can be combined. For comparison of PPML techniques, they considered effectiveness (e.g., accuracy), efficiency (e.g., running time), and privacy.

Tran et al. [69] compare different approaches such as HE, DP, SMPC and FL for privacy preserving decentralized deep learning and evaluate different artifacts such as existence of bottlenecks, privacy, utility reduction, training latency or performance costs.

Xu et al. [77] provided a framework explaining at which level different PPML techniques can be combined. This work may be considered as a combination of [81], [66], and [11]. For the first category, PPML systems are divided by phases: data preparation, model training, and model serving (deployment and inference). For the second category, PPML systems are divided according to privacy guarantee: object-oriented (data and models) and pipeline-oriented (boundary and trust assumption). For the last category, PPML systems are divided according to technical utility: data publishing approaches (elimination-based, perturbation-based, and confusion-based), data processing approaches (additive mask, garbled circuits, modern cryptographic, mixed-protocol, and trusted execution environment), architecture approaches (delegation-based ML, distributed selective SGD, federated learning, and knowledge transfer), and hybrid approaches.

Löbner et al. [45] propose a mapping to translate UAC into PPML Characteristics by eliciting influencing relationships through joint coding and expert interviews. However, they do not present how to calculate the technology that suits an application best.

To the best of our knowledge there exists no research explaining how user acceptance criteria can be used to elicit which PPML technology fits best to an application.

3 Methodology

This section presents involved entities and the methodology addressing our research objectives.

3.1 Entities Involved

We build on the entities that were already introduced by Löbner et al. [45]. Therefore, we only provide a brief summary of our definitions.

Data entity: A natural person who is providing private data for training a PPML model. No computation results are provided to the *data entity*.

User: A natural person using an AI service as a customer, providing private input data and receiving private results. *User* and *data entity* are data subjects as defined in the GDPR [27].

Developer: Commissioned by the AI service provider to create PPML models for users, using data from *data entities* and *user*. Developing a PPML model is not restricted to a single occurrence but can be done continuously. The required data to train a model is taken (a) from the data entity to e.g. built a pre-trained model or (b) from the user input.

AI service provider: Utilises the PPML model that is commissioned and build by the *developer* in a user relationship. The *AI service provider* takes over all communication with *user* and *data entity*.

Experts: Exhibit advanced background knowledge in PPML or user privacy and are important to keep the presented framework up to date.

3.2 Research Objectives

To support AI service providers and developers, we aim to identify which PPML technique a user would prefer to be implemented for an ML service that requires the disclosure of personal data. Therefore, we focus in this line of research on providing a process to evaluate different PPML techniques based on the user preferences elicited by Löbner et al. [45]. Because all PPML techniques can improve privacy but have different advantages and disadvantages a distinction is required by analysing technical differences. However, due to continuous technical improvement of PPML technologies we focus on providing a process on how to build an overall evaluation of eligible techniques (see section 4). Thus our framework is robust against changes in technology in future. We showcase the framework's application in section 6 taking over the role of framework users, utilising existing research and prototypes on the PPML technologies eligible for the use case. Finally, our approach aims to provide a

structured decision support for developers and service providers to identify the PPML techniques with good (expected) user acceptance. Thus, for this line in research we have split the ranking process of PPML techniques into three research objectives (RO):

RO1: Providing a process on how to connect PPML technologies with UAC We explain and showcase how UAC preference scores can be translated into PPML Characteristic scores. This is done by utilising the mapping provided in Löbner et al. [45]. This mapping contains already expert validated connections between UAC and PPML Characteristics. While the survey can also be conducted by the AI service provider, the translation into the PPML Characteristic scores should happen within the AI developer to adjust the mapping to the use case, if necessary.

RO2: Providing a process for evaluating the PPML characteristics that separate PPML techniques We provide a process on how the PPML Characteristics that separate different PPML techniques can be evaluated, introducing PPML criteria and criteria weights. We also showcase how the PPML criteria can be set up. This fine-grained evaluation is expected to be performed within the AI service chain, e.g., as paid service of an experienced PPML developer or as paid consultancy.

RO3: Providing a ranking method to evaluate the best PPML technique for the use case, based on RO1 and RO2 PPML Characteristic preference scores (RO1) elicited from the user input and the weighted PPML Characteristics (RO2) are used to compute a PPML technology score that can be used for technology decision support. The unique contribution of this score is that it takes user preferences into account. This task can also be performed as a service within the AI service chain because background knowledge in PPML from e.g. similar projects is likely to speed this process and increase the accuracy of the evaluation.

3.3 Threat Model

To evaluate the level of protection of the PPML techniques, it is crucial to define an individual threat model for the application in focus. Since the threat model relies heavily on the ML application, i.e. its data input and output and the aim of the application, we cannot provide a specific threat model for each possible application. To still allow our framework to consider attacks, we build a somewhat generic threat model by considering a set of common attacks that are structured based on Li et al. [40] in local and global privacy. A model protected against third parties except members of the AI service chain is considered globally private, (*resilience against attacks*) and it is locally private if protected against all entities, including internal adversaries (*purpose and access limitation*). We assume that attackers who are not part of the AI service chain, do not possess any information about the data or the model and have no physical access to the infrastructure. They can only interact with the target model by providing inputs and receiving a predicted result.

We assume that attackers inside the AI service chain, are able to collect information about the data and model over time if no protection mechanism prevents this. They might possess additional information about the user or data entity. The attacker has physical access to the infrastructure. The attackers have full access to the target model and its prediction if not prevent by PPML.

4 Framework Description

In this section, we explain the calculation steps within our framework. Figure 1 provides an overview of the different steps and clarifies at which state input is required to the framework.

First, the framework is set up by checking whether the mapping of UAC and PPML Characteristics is up to date and applicable to the application (see section 4.1). Second, User Input (see section 4.2) and Developer Input (see section 4.3) are collected. The user input is collected by conducting a survey. We assume that the developer input is a paid service by the PPML developer. It is assessed which PPML techniques for the application are eligible (see section 4.3.1). For each PPML Characteristics, categories (see section 4.3.2) with Intra-categorical weights (see section 4.3.3) are set by the developers and the PPML technologies are evaluated based on the categories (see section 4.3.4). The user input is translated by applying the mapping of Löbner et al. [45] into PPML Characteristic scores (see section 4.4). Finally, the PPML technique scores are calculated (see section 4.5). The sections of the description (section 4) correspond to the framework application (section 6). The AI service provider can now include the preference of the users for the new service into the privacy by design decisions of the application development.

4.1 Expert Input and Framework Setup

In this step, we define the framework setup using the UAC PPML Characteristics mapping proposed by Löbner et al. [45] (see table 1). Their definitions of UAC are presented for the sake of completeness in table 2. Löbner et al. [45], collected UAC and PPML Characteristics from existing models and grouped the attributes in several rounds of discussions with experts in privacy and machine learning. Moreover, experts validated the attributes and their influence onto each other in several feedback loops. Thus, we consider the framework setup as provided from previous work. How the proposed framework is applied to calculate the best PPML is the contribution of this paper. Thus, to later map the input of UAC to PPML Characteristics we define F as a binary mask matrix. A connection in the framework is represented by $F_{ij} \in \{0,1\}$, where F_{ij} denotes the connection of the j -th UAC and the i -th PPML Characteristic. The starting point are 15 UAC and 14 PPML Characteristics. For easy iteration and identification in the mapping matrix we define index j ($J = 14$) and index i ($I = 13$).

4.2 User Input

In this step, we describe the user input. The user input contains a score for every UAC. The higher the score the more important is the UAC. We define the user input u as a vector of UAC preference scores with $u_j \in [0,1]$ as the preference score for every UAC, where $\sum_{j=0}^J u_j = 1$. To evaluate the preference scores u_j a variety of possible methods exists and the choice for a method depends on the specific use case and thus cannot be part of this theoretical framework paper.

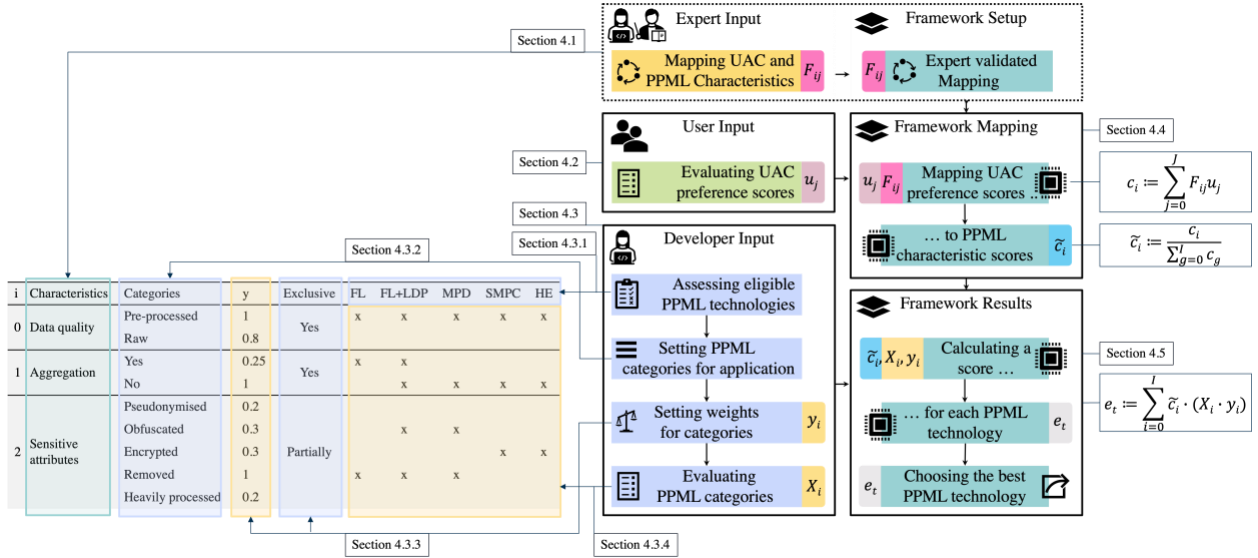


Figure 1: Flowchart illustrating the procedural steps for implementing the framework. The sections of the description (section 4) correspond to the framework application (section 6).

Table 1: Mapping of User Acceptance Criteria and PPML Characteristics

Structured Evaluation and User Acceptance Ranking for Privacy-Preserving ML Techniques

User Acceptance Criteria			PPML Characteristics														
			Data					Model						P & S			
			1. Location of data storage	2. Data size	3. Data quality	4. Aggregation	5. Sensitive attributes	1. Explainability	2. Location of computation	3. Training method	4. Accuracy	5. Training Time	6. Performance	1. Resilience against attacks	2. Purpose and access limit.	3. Technical robustness	
User Acceptance Criteria	Privacy concerns	PC1. Automated decision making	+		+		+	+			+						
		PC2. Unauthorized secondary use	+		+		+	+			+						
		PC3. Data bias			+		+										
		PC4. Unauthorized access	+				+				+			+	+	+	
	User experience	UX1. Ease of use			+						+						
		UX2. Adaptability									+						
		UX3. Availability									+						
		UX4. Performance	+	+	+				+	+	+		+			+	
	Perc. data process.	DP1. Collection	+	+		+	+				+						
		DP2. Data purpose	+	+		+	+			+							
		DP3. Storage location	+			+	+			+							
		DP4. Correctness of stored data	+		+		+										
	Perc. trustw.	PT1. Perceived lawfulness	+			+	+		+					+	+	+	
		PT2. Fairness			+		+	+		+	+						
		PT3. Transparency			+		+	+	+		+						

P & S - Privacy and Security

⬆ - User ⬆+ - User and Data Entity

Possible methods to conduct such a survey are e.g., Analytical Hierarchical Process (AHP) [70], Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [6] or the Kemeny–Young method [23].

4.3 Developer Input

In this step, we explain how the developer's input is derived. We assume that this step is a paid service performed by framework users (developers) with expert knowledge in PPML.

4.3.1 Assessing eligible PPML techniques

To identify which PPML techniques are taken into consideration for implementation, a list of possible methods can be found, e.g., in the international standard ISO/IEC 20889 [32]. This document already classifies different privacy

PC1	Automated decision making	Concern that the process is getting out of hand and people are being treated as numbers rather than individuals [64]	
PC2	Unauthorized secondary use	Concern of a misuse of data initially collected for a certain purpose, for a secondary - purpose without authorization [64]	
PC3	Data bias	Concern of discriminating results by unrepresentative data [24]	-
PC4	Unauthorized access	Concern of access of personal data to unauthorized people [64, 24]	-
<hr/>			
UX1	Ease of use	Expected effort associated with the use of a PPML technique [71, 51]	+
UX2	Adaptability	Concern that a system cannot be adapted to a change in context [72]	+
UX3	Availability	Level to which a user can successfully access a certain technology [73]	+
UX4	Performance	Extent to which the technology's use will benefit certain activities [71, 51]	+
<hr/>			
DP1	Collection	Concern that massive amounts of personal data are collected and stored [64, 80]	*
DP2	Data purpose	Has to be legitimate, explicit and specified [27]	*
DP3	Storage location	Transfer of data across geographical borders might reveal personal data, Article 44 ff. [27], Physical storage location can be local or with a cloud provider who has full data control and can perform malicious tasks [56]	*

DP4	Correctness of stored data	Concern that the stored data exhibits errors or incorrect user data [64]	*
PT1	Perceived lawfulness	Concern of a user that data is not processed lawfully [27]	+
PT2	Fairness	fairness is a substantial balancing of the involved parties to mitigate situations of unfair imbalances, where the data subject feels vulnerable. Malgieri [49]	+
PT3	Transparency	The right to receive meaningful information about the logic when automated+ decision-making or profiling is used [27]	+
UAC	Definition		*

Table 2: UAC and their expected influence positive (+), negative (-) or scenario dependent (*), from [45].

enhancing data de-identification techniques. In addition to this list, relevant literature of most recent PPML techniques should be taken into account, to cover also new emerging techniques.

4.3.2 Setting PPML categories for application

A crucial step is defining categories, denoted as k , for each PPML Characteristic. These categories should be selected to effectively distinguish between different PPML techniques. A PPML Characteristic can have any number of categories, denoted as $k \in [0, \infty)$. Categories can take on various forms, e.g. nominal, ordinal, or metric.

If a PPML Characteristic has only one valid category for a specific use case, we define it as a hard criterion. In such cases, if the hard criterion is not fulfilled, the technique can be directly excluded.

Let X_i be the matrix for the i -th PPML Characteristic, with dimensions (m, n) , where m represents the number of PPML Characteristics ($m = I$) and n represents the number of PPML techniques ($n = T$). The matrix $X_{i,kt}$ corresponds to the i -th PPML Characteristic submatrix with k categories and t PPML techniques.

4.3.3 Setting weights for categories

To cover that categories can be nominal, ordinal or metric we need to weight the categories against each other. Thus we introduce a weight vector y_i as the i -th y vector of matrix X_i . Thereby, $y_{i,k}$ is the k -th category weight for the i -th PPML Characteristic with $y_{i,k} \in [0, 1]$. Consider the following case differentiation for $y_{i,k}$:

$$\forall i : \begin{cases} \sum_{k=0}^K y_{i,k} \geq 1 & \text{if exclusive,} \\ \sum_{k=0}^K y_{i,k} = 1 & \text{if not exclusive.} \end{cases}$$

4.3.4 Evaluating PPML categories

Once the categories and weights are set up, the categories can be evaluated. The evaluation is conducted by the framework users (developers) and is expected to be performed based on previous knowledge about the respective techniques. If a technique is new, it will need to be assessed based on related literature.

If a PPML technique is assigned to a category we set $X_{i,kt} = 1$, otherwise 0. In general, a technique can be assigned to several categories, resulting in a higher score.

In the evaluation, trade-offs can be discovered. We indicate a trade-off between categories with a capital T. It is possible to set a different value below 1 making the matrix non-binary. This can be required because otherwise the result is distorted since the evaluation of the technique cannot be realised in the final implementation. The actual value for T depends on the trade-off.

4.4 Framework Mapping

In this step, we describe how the user input is translated into PPML Characteristics' scores. This is necessary because the user has a significant knowledge gap about PPML Characteristics and cannot evaluate them directly

[45]. Thus we define c_i as the PPML Characteristic preference vector with index i . It can be interpreted as the translation of UAC preference scores into preference scores for PPML Characteristics. Currently we only have the binary mask matrix F_{ij} .

Next, we calculate the PPML Characteristic scores c_i that is the dot product of the binary mask matrix F_{ij} and the UAC preference scores u_j .

$$c_i := \sum_{j=0}^J F_{ij} u_j.$$

To have a better comparison of the the results, we normalise the vector c . This is also required for the next calculation because these user preference scores for PPML Characteristics are later used to weight and rank the PPML techniques. Thus we calculate

$$\forall i : \tilde{c}_i := \frac{c_i}{\sum_{g=0}^I c_g}.$$

4.5 Framework Results

In this step, we calculate the final ranking of PPML techniques. Once all inputs are ready, the evaluation vector e can be calculated. Each PPML technique will receive a score. If other criteria besides UAC and PPML Characteristics should be included in the ranking, this has to be taken into account by the developers. In our framework, the PPML technique with the highest score will meet the UAC preference score the best. We calculate the evaluation vector e so that

$$\forall t : e_t := \sum_{i=0}^{I=13} \tilde{c}_i \cdot (X_i^T \cdot y_i).$$

5 Use Case: PSI Detection in texts

To exemplify the use of the proposed framework, we describe a simplified AI application that a provider could be planning to develop. We will use the example of PSI detection.

5.1 Use Case description

The application's objective is to detect the disclosure of Privacy Sensitive Information (PSI) in social media posts. The application should detect whether PSI was included in the text and classify the text accordingly. If the text was classified as including sensitive information, then the application would additionally detect the type of PSI contained in the text.

Different PPML methods could be used to implement such a PSI detection application. PSI detection can be accomplished using multi-class or multi-label classification approaches that would classify private data into different categories (e.g., tracking, financial or medical data) [46]. Users' privacy concerns towards this type of privacy-preserving application have been investigated by Bracamonte et al. [12], who found that users have privacy concerns with regard to perceived surveillance and perceived intrusion and secondary use, which motivates the need of PPML in this type of application. Figure 2 shows a possible interface of such a tool.

5.2 PSI detection requirements

Because of the stake of showcasing the framework, we have to set some implementation requirements and assumptions to narrow down the use case:

- The user input that is a ranking of UAC is assumed to be already achieved by a survey. How the survey is conducted precisely is part of our future experiments.
- For our example, we analyse a text reinforcement learning problem for PSI detection.

- The data used for training contains private data that cannot be removed.

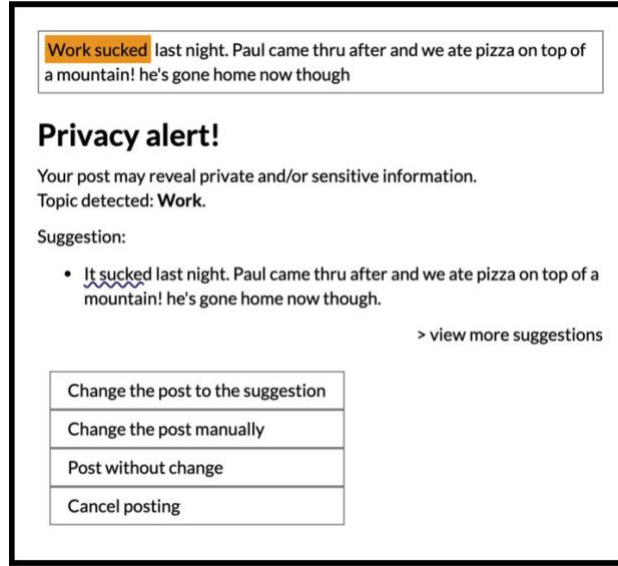


Figure 2: Example of a privacy tool for PSI detection in social media texts. [12]

- A pre-trained model based on data from the data entity already exists. We do not analyse the protection during that pre-training phase.
- PPML should be applied only to the PSI detection of user data consisting only of the input and output data but not training data provided by the data entity.
- Input data is used for improving the model. This data needs also to be protected.
- The accuracy metrics of the PPML algorithm should be as high as possible to guarantee the privacy by design principle of full functionality.
- We assume the developers using our framework to be honest and unbiased.
- If possible we want to use BERT as ML technique in combination with the respective PPML technique because it has shown very good results for text classification tasks [34, 46].

5.3 PPML for text classification

In this section, we collect examples of text classification tasks using PPML, also highlighting some limitations and advantages. This is relevant to ensure that the PPML technique is applicable.

FL is benchmarked by Lin et al. [41] who use the 20news dataset and 100 clients. They identify future potential for improved performance. Moreover, incorrect or biased data might cause biased decisions. They evaluate the current state of FedNLP as “relatively immature”. Sidhpura et al. [63] and Thapa et al. [67] propose a SMS spam and ham classification. This task is very similar to a binary PSI detection because the length of text is comparable. Liu et al. [42] analyse FL for Natural Language Processing (NLP) and find relevant challenges for text classification, e.g. solving communication overhead from client updates using distillation methods. A benchmarking of FL with NLP is provided by Lin et al. [41] who analyse a combination of NLP with BERT and FL.

FL+LDP is combined by Löbner et al. [44] to increase the privacy of users. For a noise multiplier of 0.99 and an ϵ of 23 they achieve an F1-Score of 0.94 after 20 federated rounds for 10 clients. In this approach LDP is applied on the gradients of the users’ local models and not on the text itself. The evaluation of the whole model took 144 min.

DP for unstructured text data was investigated by Klymenko et al. [35]. They find that DP does not permit inference attacks themselves but creates uncertainty about the inferred data. Regarding unstructured text, this concept can run into limitations when following the strict definition that any two texts have to be considered adjacent. Thus, the application of Metric Differential Privacy (MDP) was introduced by Chatzikokolakis et al. [19] for unstructured text

data. Carvalho et al. [17] apply MDP on the IMDB review data set, using a sentiment classification model. Xu et al. [79] present a differentially private Mahalanobis mechanism for text perturbation. Weggenmann and Kerschbaum [75] propose an automated text anonymisation approach for text mining that fulfills DP and comes with a provable plausible deniability guarantee. Classifying text into news groups, they find that their model requires a large ϵ to achieve an acceptable utility which has negative privacy consequences. The application of LDP for unstructured text classification is possible but comes with some drawbacks such as unclear definitions and privacy guarantees. Thus we decided to analyse MDP instead of LDP. Also the privacy accuracy trade-off has to be considered carefully.

SMPC is used by Resende et al. [58] for secure classification of spam and ham emails in a two-party protocol. They notice no loss in accuracy by applying SMPC with Naive Bayes and can compute the result in 21ms for an average spam SMS of 8 unigrams. They state that in their model the service provider cannot read the text and the user does not learn anything about the model, except the classification result. Knott et al. [36] present the CryptTen protocol that utilises a linear layer operating on word embedding. Compared to PyTorch they noticed a reduced speed (about 2.5-3 orders of magnitude slower) with an overall inference per sample of 30ms. Comparing a two- and three-party protocol they find that the latter is slower, due to the larger number of communication rounds required by the public division protocol. Reich et al. [57] use SMPC for hate-speech detection in personal text messages. They also prove that with their classification approach based on feature extraction with logistic regression and tree ensembles the users cannot learn about the model and the service provider does not learn anything about the user's message. Based on the collected examples we conclude that SMPC is applicable in our PSI detection application, even with NN.

HE is implemented by Al Badawi et al. [3] who use fully encrypted data to train an encrypted model for spam and ham classification. They achieve a run time per inference of 170ms. They have no loss on prediction accuracy compared to the baseline model without a privacy layer. Moreover, they show how inference as a service can be implemented by a service provider, ensuring privacy of the data and privacy of the model. Podschwadt and Takabi [55] present a RNN for NLP tasks without a loss in accuracy. In their example they perform a sentiment classification on movie reviews with a runtime of 5800 ms for a batch size of 32. Sun et al. [65] implement a FHE scheme with SIMD that performs private hyperplane decision based, Naive Base and decision tree comparisons for cyphertexts. They achieve an implementation time for Naive Bayes between 49 and 142 ms for different security parameters (λ). Thus HE is suitable for our application but is compared to SMPC a little bit slower. Also a trade-off between security and performance exists.

TEEs are secure and isolated processing environments, distinct from the regular processing environment. By partitioning applications and running sensitive code in the TEE the security and data integrity are increased [26]. Sartakov et al. [61] describe the use of TEE in clouds to deploy applications that are protected from e.g. unauthorised access by cloud service providers. Using lift-and-shift models whole VMs can be run in TEEs. Thus we consider this technology.

6 Framework Application

In this section, we apply the framework as described in section 4 for the PSI detection use case as presented in section 5. In the role of framework users, we will go through each step of the framework and calculate the PPML technology that meets the user requirements the best.

6.1 PSI: Expert Input and Framework Setup

We do not implement any changes to the framework itself and consider all UAC and PPML Characteristics to be relevant for the PPML use case. We do not include additional features. Experts should re-evaluate the framework at regular intervals based on technology changes.

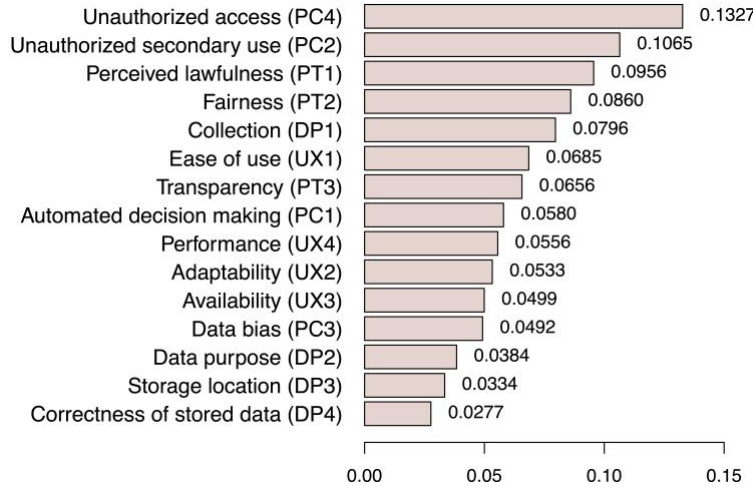
6.2 PSI: User Input

For the ranking we have run an AHP with 55 participants. Due to the high number of UAC (15), the UAC were divided into four groups (PC, UX, DP, PT) and the respective subgroups as provided in table 1. The AHP took each

participant around 15 minutes and a tolerable [31] Consistency Ratio (CR) below 0.2 was achieved for 31-33 participants in each subgroup. Participants above 0.2 were excluded. Between 18 and 21 participants achieved a CR score below 0.1. 27 participants were male and 27 female, 1 participant was diverse. Four participants were in the age group 18-19, 42 between 20-29, seven between 30-39, none between 40-49, one between 50-59 and one between 60-69. We aimed specifically for a younger age to get people with high social media interaction. The usage and posting behavior is shown in table 3. Following [60] when using AHP for group decisions we expect a relative importance score for each feature between 1 and 0. All global feature preference values sum up to 1. We also follow the fundamental scale: {extremely important : 9, very strong important : 7, strong important 5, moderate important : 3, equal : 1 }. We have also included intermediate choices (2, 4, 6, 8). The global preference values are provided in figure 3.

Quantity	How often do you use social media?	How often do you post in social media?
Daily	33	4
Weekly	5	11
Monthly	11	22
Yearly	1	12
Not at all	5	6

Table 3: Social media usage and posting behavior


 Figure 3: Importance scores u_j (x-axis) of UAC (y-axis)

6.3 PSI: Developer Input

In this section, we describe how we derived the developer input for the PSI detection application.

6.3.1 PSI: Assessing eligible PPML techniques

Techniques are selected based on related literature, ISO/IEC 20889 and our own experience. We assume that also developers would choose techniques based on their own experiences and the techniques establishment. As shown in section 5.3 all techniques chosen for showcasing are suitable for the PSI detection application:

- Federated Learning (FL)
- FL + Local Differential Privacy (FL+LDP)
- Metric Differential Privacy (MDP)
- Secure Multiparty Computation (SMPC)
- Homomorphic Encryption (HE)
- Central Trusted Execution Environment (TEE)

6.3.2 PSI: Setting PPML categories for application

In this section, we elicit categories for the PPML Characteristics based on table 1. An initial definition of the PPML Characteristics is already provided by Löbner et al. [45]. We have separated the PPML Characteristics into categorical attributes that do not have a logical order, but one category might be preferred by the users.

For a better overview and easier computation of the final recommendation, we start by defining hard requirements that will lead to an exclusion of the PPML if not fulfilled. We do this for an improved efficiency of the evaluation process. The following criteria of PPML Characteristics we define as a hard requirement.

Location of raw data storage: The location where raw user data is stored is for our PPML techniques in focus local (stored on a user’s device), central (stored on a central server within the AI service chain) or distributed (stored across several entities). For our PSI application we set the requirement to only allow locally stored raw user data. We set this restriction because survey participants clearly preferred raw data to be stored exclusively local.

Table 4: Hard requirements for exclusion of technologies including PPML Characteristics and categories (see section 6.3.2), and PPML technique evaluation (see section 6.3.4).

Characteris.	Categories	FL	FL+LDP	MDP	SMPC	HE	TEE
Location of raw data storage	Local	x	x	x	x	x	
	Distributed						
	Central						x
Data size	Min. 1k sampl.	x	x	x	x	x	x
Training method	Supervised						
	Semi-supervised						
	Unsupervised						
	Reinforcement	x	x	x	x	x	x

Data size: Since for some ML models, a huge data size is mandatory, a threshold should be set for the data size that indicates the minimum amount of data, e.g. in X (formally known as Twitter) posts, required to achieve a useful accuracy. In the papers investigated by Löbner et al. [46] the amount of text samples in the data set was between 1,000 and 800,000. Thus, we will just provide the threshold of minimum 1,000 samples to indicate that for the application pre-trained models have to be used. Since we only consider short text posts on social media, we neglect the text-length.

Training method: Not all PPML techniques are suitable for each training method. Frequently used training methods are e.g., *supervised learning*, *semi-supervised learning*, *unsupervised learning* or *reinforcement learning* [45]. Thus we will evaluate for common tasks whether our PPMLs in focus for the PSI application are suitable for a NLP task that is part of reinforcement learning.

Next, we evaluate the hard requirements for all PPML Characteristics (see table 4).

Regarding the *location of data storage*, all technologies except a central TEE allow to store raw data exclusively local. We characterise raw data as the unaltered user’s text that contains potential private information. FL, FL+LDP, SMPC and HE alter the data locally, e.g. on a mobile phone before sharing with other entities [41, 44, 17, 58, 3]. With a central TEE the data is required to be send to a trusted entity running the TEE. The PPML Characteristics *data size* and *training method* are fulfilled by all PPML technologies since comparable application were found in section 5.

Next, we explain how we set up the categorical scales for the different PPML Characteristics that are not hard requirements. All PPML Characteristics that are not hard requirements are provided with their respective categories in table 5. It is important to keep in mind that for other applications and especially other PPML techniques, the scales might require an extension.

Data quality: First, data quality is influenced by outliers and missing data that cannot be handled well by all algorithms. Second, data quality is influenced by the process of removing, aggregating, obfuscating, or changing data for deidentification. Thus there is an inverse relation with the attribute *aggregation*. Regarding text data for PSI detection, it is known that the data has to be pre-processed since texts from social media are full of shortened words, abbreviations or colloquial language that complicate the analysis [46].

Aggregation: According to ISO/IEC 20889 aggregated data is representing statistically the attributes of several data entities and is only useful in specific use cases. Aggregation has an inverse relation with *data quality* and *accuracy*. We indicate with *yes* or *no* whether *aggregation* is happening on the raw data during the de-identification.

Sensitive attributes: Plain sensitive data is one of the most important issues in PPML. To prevent the revealing of plain sensitive data we differentiate between the following fundamental methods of *pseudonymisation*, *obfuscation*, *encryption*, *removal* and *heavy processing*. *Heavily processed data* we define as processed in a way that the original private data is no longer recognisable. Although removing sensitive data is the best method to break re-identification attacks, this can be rarely done if accuracy is required to be high.

Explainability: Explainable results help to deploy proper PPML techniques [77]. There exist *ante-hoc* methods that are explainable by default such as decision trees, *post-hoc* methods that are methods applied after the model was built [16], *global methods* that explain the overall general behaviour of the model [13] and *local methods* that aim to explain individual, specific decisions [1].

Location of computation: For simplicity reasons, we distinguish between three possible locations of computation. These are *cloud* computation, computation on *local* user devices and *distributed* computation. Computation refers to the training of the model [45].

Accuracy: An accuracy threshold for the suitable accuracy metrics should be defined. If a model in combination with the PPML is not able to achieve the threshold, the PPML should not be further considered. From Löbner et al. [46] we know that for a binary classification of PSI data into privacy sensitive and non-sensitive texts, related models achieve an F1-score of 0.98 [37]. Thus, for the binary classification we decided for the categories $< 0.84, 0.85 - 0.89, 0.9 - 0.94, 0.95 - 1.00$.

Training time: Training time influences the adaptability of a PPML model and depends on the computational complexity of the training task [45]. Especially for reinforcement learning problems large amounts of training time are required which are often compensated with the use of GPUs or FPGAs [59]. Training time is very hard to compare between different approaches because complexity of models, data sets, computational resources and parties involved differ. Nevertheless, we want to elicit from existing models first insights on the suitability for PSI detection. To cover this issue we set a very general scale of categories $< 12h, 12h-24h, > 24h$.

Performance: Performance is the overall runtime of an application. Regarding the PSI classification task, it is the time between user request and receiving the classification result. Egger et al. [25] elicited three thresholds from literature. First, 0.1 s is the threshold for feeling an instant reaction of the system. Second, 1.0 s is the threshold a delay is noticed but the flow of a user's thoughts is not interrupted. Third, 10 s is the threshold of keeping the attention on the dialogue between application and user. We follow these three limits for the evaluation of PPML techniques.

Resilience against attacks: As explained in section 3.3, we consider the definition of "globally private" which refers to attackers external to the AI service chain. We consider only attacks which leak information to the external adversary, i.e. *model inversion attacks* [30] and *membership inference attack* [62]. Our characteristic reflects if the considered PPML approach is resilient against this kind of attack. There also other attacks such as poisoning attacks [8] or evasion attacks [9] which do not aim to infer data but to manipulate the results of the model's answer. Since these attacks are more related to the underlying ML approach we consider them under model robustness.

Purpose and access limitation: As already explained in section 3.3, for this characteristic we consider the definition of "locally private" which refers to attackers inside the AI service chain. Based on Tanuwidjaja et al. [66], we differentiate between privacy of user, privacy of model and privacy of result. Due to our distinction between users of the system and data entities who only provide training data for the system, we expanded their classification by adding data entities. Contrary to the external attackers from the previous characteristic, each member of the AI service chain needs "some access" to the data in order to fulfil their task. Therefore, for this characteristic, we evaluate whether the PPML methods implements countermeasures to protect each of these data layers to ensure members of the AI service chain do not use the data for other purposes than intended. E.g., blackbox reconstruction attacks are considered here for entities that can access the results. Again, we only focus on how the PPML method can help to improve the privacy, thus concepts such as authorisation management are not considered.

Technical robustness: According to the Ethic Guidelines for Trustworthy AI [22] technical robustness should ensure that even with small changes in the operating environment or the context of the model the PPML behaves reliable

during the whole AI life cycle, minimising unexpected or unintentional behaviour. In this framework we focus on the evaluation of the susceptibility to adversarial attacks based on existing literature, especially model poisoning [21]. In model poisoning attacks the adversary tries to change the model's prediction by polluting the model with manipulated input data. Thereby, the attack can be targeted to achieve a certain classification result or untargeted, just polluting the model and lowering its accuracy [68].

6.3.3 PSI: Setting weights for categories

For each category matrix X_i we have to set a weight vector y_i . The sake of these weights is to support framework users with another dimension to formulate requirements. Again the sum of each weight vector if not exclusive has to add up to 1. Setting up the weight vector is a task that is performed in the AI service chain by the party who has the most knowledge about the PPML characteristics. To achieve an equal scaling, an AHP could be used again for each category. All values for y are shown in table 5 where we also indicate whether a category is exclusive or not. If a category is not exclusive ("No"), then $\sum y = 1$ holds. For the sake of showcasing, we will set the weight vectors exemplary based on authors' experience (see table 6)

6.3.4 PSI: Evaluating PPML categories

In this section we describe the evaluation of PPML technologies (FL, FL+LDP, MDP, SMPC, HE) by labelling them into PPML categories from section 6.3.2. For the sake of showcasing, the authors take over the role of framework users. To evaluate the technologies we use the first 50 papers in ACM, IEEE and Google Scholar that were identified by search terms that consist of the respective technology, PPML Characteristic respective PPML criteria. The results of the evaluation are presented in table 7.

Table 5: PPML Characteristics with categories (see section 6.3.2) and category weights y (see section 6.3.3).

i CharacteristicsCategories		y	Exclusive
1	Pre-processed	1	Yes
	Raw	0.8	
	Aggregation Yes	0.25	No
	No	1	

		Yes
Pseudonymised	0.2	
Obfuscated	0.3	
Encrypted	0.3	Partially
Removed	1	
Heavily processed	0.2	
Ante-hoc methods	0.25	
Post-hoc methods	0.25	
Global explainability	0.25	No
Local explainability	0.25	
Local	0.5	
Distributed	0.3	No
Central	0.2	
<0.84	0	
0.85-0.89	0.3	
0.9-0.94	0.6	Yes
0.95-1	1	
Attention lost: >10 s	0	
Attention kept: 1.0-10 s	0.3	
Uninterrupted: 0.1-1.0 s	0.6	Yes
Instant reaction: <0.1 s	1	
Membership inference	0.5	No
Privacy of data entity	0.25	

 Table 6: Attributes, Justifications, and y Values

PPML Characteristic	Justification	y
Data Quality	The effort to create pre-processed text data from raw is low because a lot of packages exist already and the BERT model has its own pre-processing layer.	[1,0.8]
Aggregation	Can cause information loss, thus no aggregation is preferred.	[0.25,1]
Sensitive Attributes	Handled differently, with the removal having the highest protection, and obfuscation and encryption evaluated as effective but vulnerable to attacks.	[0.2,0.3,0.3,1,0.2]
Explainability	The more explainability the better.	[0.25,0.25,0.25,0.25]
Location of Computation	Data not leaving the device cannot be sold to third parties or accidentally altered. This also holds for distributed setups although the privacy guarantees are often lower compared to local computation.	[0.5,0.3,0.2]
Accuracy	The higher the better but below 0.84 is not acceptable.	[0,0.3,0.6,1]
Training Time	Over 24h quick adaptation to environmental changes is prevented.	[0.3,0.6,1]
Performance	Faster is better but losing attention is not acceptable.	[0,0.3,0.6,1]
Res. Against Attacks	All attacks should be prevented.	[0.5,0.5]
Purpose and Access Lim.	Should be equally implemented.	[0.25,0.25,0.25,0.25]
Technical Robustness	Important over the whole PPML life cycle; data poisoning needs to be prevented.	[1,0]

Data quality: To achieve a high accuracy in the PSI detection task, pre-processing is required for NLP. All PPML are compatible with our requirement to work with BERT [5, 2, 38] that requires pre-processing. The pre-processing usually happens in a pre-processing layer of the BERT model.

Aggregation: A typical task within FL is to aggregate the gradients of the clients with an aggregation function to create an updated model [52]. Thus aggregation is automatically happening for FL and FL+DP. The updated model might, due to aggregation, not contain the best gradients for a specific client. MDP, SMPC and HE are not using aggregation before the actual ML task.

Sensitive attributes: FL is preserving privacy by the “lessening footprint” of private user data in the gradients that are submitted to the central server [52]. Thus we introduced the category *heavily processed* because in contrast to *removed* the private attributes can be reconstructed through several attacks, e.g. by a malicious central server. Regarding FL+LDP, an additional layer of privacy is added to the gradients before they are shared with the central server. In the ISO/IEC 20889 LDP is evaluated as useful for data minimisation and in scenarios where the data receiving entities cannot be trusted. LDP uses random noise from a “carefully selected” probability distribution that is added to the data to make the algorithm differentially private while simultaneously preserving the desired usefulness [32], thus we map this as *obfuscation*. We follow Tanuwidjaja et al. [66] and classify SMPC and HE as cryptographic approaches that use encryption. HE is computing on encrypted data thus the main technique here is *encryption*.

Explainability: In FL, it is possible to use all existing explainability-methods but it is important to differentiate between the explainability of the local and global model with a focus on not revealing private data. Regarding DP all explainability methods are possible Patel et al. [54], Naidu et al. [53]. Thus, the same holds for FL+LDP. Regarding explainability in SMPC and HE we did not find reliable results. Therefore, we assume that black-box methods will work and post-hoc methods and local methods are possible. Due to the encryption we determine that ante-hoc methods and global explainability are not suitable or hard to implement. These methods could reveal too much information about the model and thus leverage attacks.

Location of computation: FL as used in our scenario relies similar to Lin et al. [41] on a central server. Computation is done mainly locally but for the model updates a central model is computed with the clients’ gradients. This improved model is sent back to the clients as an updated model. Regarding FL+LDP the *locations of computation* does not change since with LDP only an additional step on the client is added that is also computed locally [29]. In contrast to LDP, we assume MDP to be applied centrally before data is computed. SMPC uses distributed protocols to derive a result, thus the computation is happening distributed [58]. HE computes results and model centrally [3].

Accuracy (F1-Score): For the problem of PSI detection we expect for a binary text classification a F1-Score of 0.97 or better [37]. Since the referred models in this section do not deal with the same data and classification problem, accuracy scores and metrics provide just a range of expected accuracy. Related work achieved for FL in a binary text classification task an F1-Score over 0.95 [67, 63]. Text classification tasks with FL and an additional LDP layer exhibit a reduced accuracy due to the privacy-utility trade-off, achieving a F1-Score between 0.85 and 0.89 [44] for a binary text classification task. Also Basu et al. [4] notice a reduction of accuracy caused by DP in comparison to their baseline model of approximately 5%. Carvalho et al. [17] achieve an average accuracy of 75% using an MDP approach. SMPC in general does not exhibit the issue of accuracy loss and thus an accuracy over 0.95 is expected [58]. Also HE has no expected loss of accuracy compared to a baseline model trained without PPML, thus we expect an F1-Score over 0.95 [38].

Training time: Xu et al. [78] report a training time for PSI detection in texts for LSTM of approximately 16 h and CNN off approximately 2 h. For FL in tests with 10 clients for a binary text classification task of spam and ham, 40 training rounds with an LSTM model took Löbner et al. [44] 137 minutes. Adding an LDP layer for 10 clients extended the training time to 144 minutes. Regarding HE studies report a training time of 5.04 days [3]. One reason identified a huge data transfer rate between GPU and CPU. Regarding MDP we assume that there is no notable delay caused to the training time. Although we did not identify any related work reporting training times for MDP in text classification, the training performance itself is expected to be increased only slightly because more epochs might be required. To the best of our knowledge, no reports about training time for privacy preserving training exist. Most studies identified [47, 58, 74, 39] rely on pre-trained models but do not train with secure multiparty computation itself. Thus we do not tick a box for SMPC which will result in a lower score compared to the other technologies.

Performance: Xu et al. [78] report for their best model a detection time of 0.21 s what we denote as uninterrupted classification. Since FL and FL+LDP rely on local models, we expect a time similar to the baseline model and thus, classify the performance as uninterrupted. Resende et al. [58] evaluate the runtime of their SMPC classification model in the worst case 0.3 s and in the best case 0.022 s and outperform Reich et al. [57] who report a total classification time between 7.2 and 13.3 s. Also Knott et al. [36] report a classification of text in 0.03 s. In average we assume an uninterrupted performance for SMPC. Al Badawi et al. [3] report the computation of a classification result within 0.3 s. Regarding HE we identified a prototype reporting a performance of approximately 6 s and thus, set *attention kept* [55]. For MDP we have not found any reports about the performance. Based on LDP performance [28] and the short text-length, we expect the performance to be uninterrupted.

Resilience against attacks: FL alone is not resilient against model inversion, adversarial attacks or membership inference attacks [52]. Adding an LDP layer, the resilience against external adversaries can be improved [44],

protecting the model from membership inference attacks. There is a trade-off between accuracy and privacy that needs to be considered. There is no sufficient protection against model inversion attacks because updated models are not protected. Also MPD is applied on the data and thus helps only with membership inference but not with model inversion attacks. Regarding SMPC, the model and the users' data are well protected against external adversaries [58]. Also HE has the potential to protect against model inversion and membership inference attacks from external adversaries. In their model that combines BERT and HE Lee et al. [38] significantly reduce the inversion risk, especially the black box inversion attacks cannot be applied using an 128-bit security level. With only the client being able to decrypt the classification result, membership inference attacks from external adversaries can be prohibited.

Purpose and access limitation: Kairouz et al. [33] find that although FL improves privacy compared to centralized training approaches no formal privacy guarantees exist. Thus information leakage between client and central server is possible. Especially over several rounds of training the central server can learn model parameters and compromise user data. Since the data from the data entity is normally used to pre-train a model by the central server, there is no protection against attackers from inside at all. Adding a differential privacy layer on top of each client helps to protect the local model. Thus privacy of model and privacy of user are increased but a trade-off exists [44]. On the one hand, it takes more federated rounds until the central server can learn about the model, on the other hand, the DP layer is very likely to reduce accuracy of the model. The DP layer at the client does not protect the data entity. MPD can be applied on the user data as well as on the data from the data entities, but it does not help to protect the model or results. Still, a privacy-accuracy trade-off exists. Regarding SMPC other participants are learning nothing about the result or the user data [58]. We assume that the protocol is protected from user access and thus privacy of the model can be achieved. Al Badawi et al. [3] show that HE can protect the input data (privacy of data entity, privacy of user), privacy of model and the classification result because each client encrypts the data before passing it to the central server [66].

Technical robustness: In this section, we evaluate whether PPML techniques are vulnerable against model poisoning and whether mitigation strategies already exist. Model poisoning is a huge problem in FL [14] but countermeasures exist already for FL that guarantee comparable performance [43, 48]. Regarding LDP protocols are in theory vulnerable to model poisoning [10, 77], e.g. Cao et al. [15] show that it is possible for an adversary to inject fake users to an LDP protocol and successfully send carefully manipulated data that is accepted by the central server. Xu et al. [76] analyse mitigation strategies for poisoning in DP in text classification problems and reduce the attack success rate from 0.94 to 0.008 while accuracy is only reduced by 0.005. Regarding SMPC [20] and HE we expect that model poisoning is also possible and needs countermeasures if the user data is also used for training. We identified a lack in literature investigation model poisoning in SMPC and HE.

6.4 Framework Mapping

In this section we describe how to map the importance scores of the PPML characteristics and the PPML category evaluation metrics. As mentioned in the assumptions we only take the user and not the data entity into account. The computed normalised \tilde{c}_i values for each PPML characteristic are provided in figure 3 for User and Data entity. The translation of UAC preferences scores into PPML Characteristic preference scores is presented in figure 4.

6.5 PSI: Framework Results

In this section, we derive the technology ranking based on the previous table by calculating a score that reflects the technology evaluation from table 7 and the PPML importance scores from figure 4a.

The first step in this calculation is to translate table 7 that represents the values of X into numbers we can use for calculation. E.g., for the PPML Characteristics *sensitive attributes* X_2 and y_2 are represented the following:

$$\begin{array}{cccccc}
 0 & 0 & 0 & 0 & 0 & 0.2 \\
 0 & 1 & 1 & 0 & 0 & 0.3 \\
 X_2 & 0 & 0 & 1 & 0 & y_2 \\
 = 0 & 0 & 0 & 0 & 1 & = 0.3 \\
 0 & 1 & 1 & 0 & 0 & 0.2 \\
 1 & & & & 0 & 1 \\
 & & & & 0 & 0.2
 \end{array}$$

The hard requirements have to be fulfilled and a technology will be removed if it does not. In a later implementation we plan to invoke a notification if a hard requirement is not fulfilled. Since we have only evaluated techniques that do fulfill the hard requirements we reduce l to 10. Regarding the trade-offs (see table 7) we have set $T = 1$ because the trade-off was already considered when setting e.g. a lower expected accuracy for MDP. We first derive the dot product of X and y (see table 8). Now we can compute the score e using our formula that includes the user preference scores \tilde{c}_i , the technology characteristic evaluations X_i and the weight vectors y_i :

Table 7: PPML Characteristic evaluation (see section 6.3.2). Evaluation (see section 6.3.4)

Characteristics	Categories	FL	FL+LDP	MDP	SMPC	HE
Data quality	Pre-processed	x	x	x	x	x
	Raw					
Aggregation	Yes	x	x			
	No			x	x	x
Sensitive attributes	Pseudonymised		x	x		
	Obfuscated					
	Encrypted					
	Removed				x	x
	Heavily processed	x	x	x		
Explainability	Ante-hoc methods	x	x	x		
	Post-hoc methods	x	x	x	(x)	(x)
	Global explainability	x	x	x		
	Local explainability	x	x	x	(x)	(x)
Location of computation	Local	x	x			
	Distributed	x	x		x	
	Central			x		x
Accuracy (F1score)	<0.84			T		
	0.85-0.89		T			
	0.9-0.94					
	0.95-1	x			x	x
Training time	>24h 12-24h				(x)	x
	<12h	x	x	(x)		
Performance	Attention lost: >10s					
	Attention kept: 1.0-10s					x
	Uninterrupted: 0.1-1.0s	x	x	(x)	x	
	Instant reaction: <0.1s					
Resilience against attacks	Model inversion				x	x
	Membership inference				x	x
Purpose and access limitation			T	T		
	Privacy of data entity			T		x
	Privacy of user		T	T	x	x
	Privacy of model		T		(x)	x
	Privacy of result				x	x
Technical robustness	Poisoning counterme.	x	x	x		
	No countermeasures				(x)	(x)

Based on literature := x, Trade-off := T, Estimate := (x)

Structured Evaluation and User Acceptance Ranking for Privacy-Preserving ML Techniques

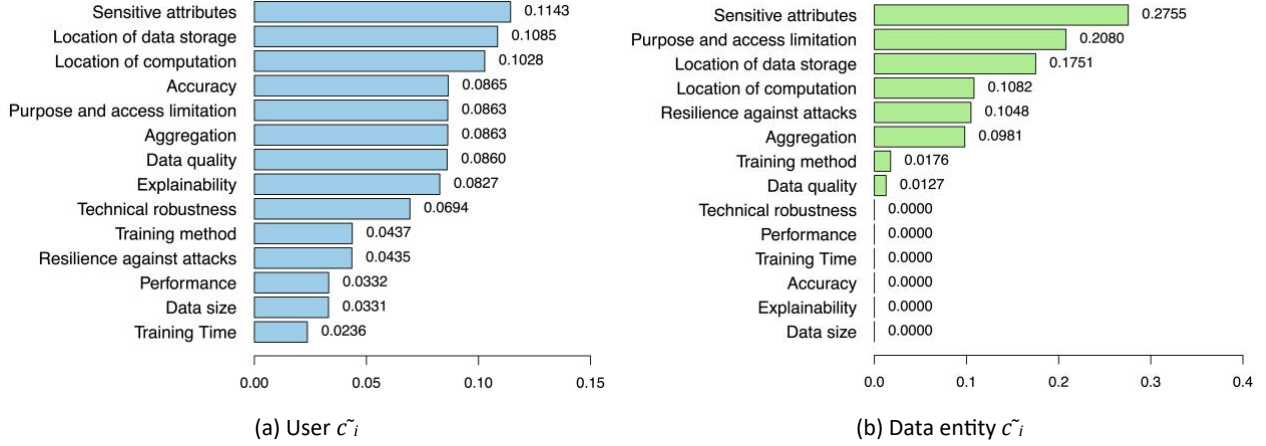


Figure 4: Importance scores \tilde{c}_i (x-axis) of PPML Characteristics (y-axis) for user/data entity

$$l=10$$

$$e := \sum_{i=0}^l \tilde{c}_i \cdot (X_i^T \cdot y_i)$$

The final scores of e are shown in table 9. In addition we also show the intermediate results for each PPML Characteristic and Technology. These results already contain the category weights (y) and the users' normalised PPML Characteristic preferences (\tilde{c}_i).

Table 8: Users' PPML Characteristic scores for each PPML technology: $(X_i \cdot y_i)$

	FL	FL+LDP	MDP	SMPC	HE
Data Quality	1.00	1.00	1.00	1.00	1.00
Aggregation	0.25	0.25	1.00	1.00	1.00
Sensitive Attributes	0.20	0.50	0.50	0.30	0.30
Explainability	1.00	1.00	1.00	0.50	0.50
Location of Computation	0.70	0.70	0.20	0.30	0.20
Accuracy F1	1.00	0.30	0.00	1.00	1.00
Training Time	1.00	1.00	1.00	0.30	0.30
Performance	0.60	0.60	0.60	0.60	0.30
Resilience Against Attacks	0.00	0.50	0.50	1.00	1.00
Purpose and Access Limitation	0.00	0.50	0.50	0.75	1.00
Technical Robustness	1.00	1.00	1.00	0.00	0.00
SUM	6.75	7.35	7.30	6.75	6.60

Table 9: Users' PPML Characteristic scores for each PPML technology: $\tilde{c}_i \cdot (X_i \cdot y_i)$

	FL	FL+LDP	MDP	SMPC	HE
--	----	--------	-----	------	----

Data quality	0.086036	0.086036	0.086036	0.086036	0.086036
Aggregation	0.021572	0.021572	0.086289	0.086289	0.086289
Sensitive attributes	0.022865	0.057162	0.057162	0.034297	0.034297
Explainability	0.082736	0.082736	0.082736	0.041368	0.041368
Location of computation	0.071946	0.071946	0.020556	0.030834	0.020556
Accuracy F1	0.086531	0.025959	0.000000	0.086531	0.086531
Training time	0.023646	0.023646	0.023646	0.007094	0.007094
Performance	0.019897	0.019897	0.019897	0.019897	0.009949
Resilience against attacks	0.000000	0.021752	0.021752	0.043504	0.043504
Purpose and access limitation	0.000000	0.043144	0.043144	0.064716	0.086289
Technical robustness	0.069414	0.069414	0.069414	0.000000	0.000000
e	0.484643	0.523264	0.510632	0.500566	0.501913

Comparing the final results we see that FL+LDP has achieved the highest score. Comparing the PPML Characteristic scores, we notice that no technology is dominating another. Comparing FL and FL+LDP the addition of LDP increases the score for *Resilience against attacks* and *Purpose and access limitations* (see table 8). The cost of this is reduced accuracy, evaluated by a lower F1-Score. This is not surprising and comes from the fact that DP in general is likely to decrease the accuracy in favour of privacy. This is also notable when comparing FL with MDP. The biggest difference in these architectures is that LDP is used locally on the gradients of the FL architecture while MDP is applied directly on the data at a central entity. The highest scores in privacy and security characteristics were achieved by SMPC and HE while HE has the best scores. Comparing the results in table 8 and table 9 that takes the user weights into account the relative importance of SMPC has slightly decreased compared to HE. The reason for this is that *performance* where SMPC is superior was assessed less important (see figure 4a) with 0.0332 compared to the other PPML characteristics. The score of FL+LDP was increased compared to the second best solution because location of *computation* has a high importance for the users with 0.1085. For completeness and the sake of showcasing we have also used the potential user preferences to compute the data entities' results that are presented in table 10. In this table, we have removed all rows that do not have an impact on PPML Characteristics based on table 1 and would thus just exhibit the value 0 (compare figure 4b). We find that the preference for FL and FL+LDP are the lowest for the data entity compared to the users. For a more comprehensive statement it might be useful to collect the preferences of data entities directly when collecting their data. This is more realistic than using data of potential users and can help to avoid bias in data.

How to weight between data entities and users and which technology is finally chosen is not trivial since further features can impact the decision, e.g. maintenance costs. Still, our framework provides detailed insights in most important PPML Characteristics that satisfy the users' needs.

7 Discussion

In this section we discuss our impact, lessons learned and future work.

Table 10: Data entities' PPML Characteristic scores for each PPML technology: $\tilde{c}_i \cdot (X_i \cdot y_i)$.

	FL	FL+LDP	MDP	SMPC	HE
Data Quality	0.012707	0.012707	0.012707	0.012707	0.012707
Aggregation	0.024529	0.024529	0.098116	0.098116	0.098116
Sensitive Attributes	0.055104	0.137760	0.137760	0.082656	0.082656
Location of Computation	0.075706	0.075706	0.021630	0.032445	0.021630
Resilience Against Attacks	0.000000	0.052423	0.052423	0.104845	0.104845
Purpose and Access Limitation	0.000000	0.103977	0.103977	0.155966	0.207955
e	0.168046	0.407102	0.426613	0.486735	0.527909

7.1 Impact

With the rise of the AI Act, the need for PPML has been confirmed by the European Commission. Incumbents that are in charge to implement privacy for their AI application as well as new entering companies can utilise our framework to understand the needs of their customers and thus adjust the privacy strategies. We aim to strengthen

the principles of privacy by design by providing clarification on which PPML technology will bring the highest user satisfaction. This aims to ensure full functionality while providing a high privacy level. Regarding research we identified a gap in structured comparison of PPML technologies that take user preferences into account. We hope to encourage other researchers to close this gap.

7.2 Empowering of user rights

With this work we empower the user of an AI service with the possibility to choose a technology that meets best their preferences. Although the users are by our approach technically empowered to choose the PPML with the best security and privacy, this does not necessarily mean that they do so. Nevertheless, optimizing efficiency of an application by trading e.g. performance against privacy is an intended possible outcome of the framework. We follow the approach that security and privacy in an application is only useful if the application still provides full functionality [18].

7.3 Lessons learned

During the evaluation of the categories we noticed that dependencies between categories exist. E.g., for DP often a trade-off between accuracy and privacy is reported. While we take this into account with a reduced weight for these dependencies, we will further investigate how to best implement trade-offs in an advanced version of the framework.

To implement user preferences for a PPML technology decision requires many steps and specific technological expertise. How to improve the evaluation time is a topic we will address in future.

Since we have experimented with a lot of different UAC rankings we noticed that small changes in the UAC preference scores can have a huge impact on the final result. Thus, robustness of the model should be investigated with further studies in future.

When analysing the different steps to derive the framework results we had several feedback loops to ensure that there is notable impact of UAC preference scores on the final results. To get a good balance between technology and UAC preference score influence is not an easy task. In our future work, we want to dive deeper into the explainability and tracing of influences.

7.4 Limitations

While the framework is a first step towards the possibility to consider user preferences for the selection of a PPML technology, it requires parameters specific to each considered scenario. For example, some of the thresholds and weights for PPML Characteristics (cf. Table 7) might be different for another scenario. Determining these parameters can take a considerable amount of effort and requires expertise in several disciplines.

Regarding the application of the framework in the PSI detection use case, we had to rely on the statements from existing literature, e.g., regarding privacy guarantees or run-times of the different technologies. It is important to be aware of the setup differences in the literature and resulting reported accuracy and performance disparities.

The values in table 6 are an example and might be influenced by personal preferences since it can be hard to come up with reasonable values and scales, especially if literature is scarce.

Another limitation is the number of participants in the AHP that we surveyed to calculate the UAC preference scores. However, since we were only interested in working through the use case as a demonstration of our framework, the impact is negligible.

7.5 Future Work

Since we have just used a small group of people participating in the AHP we will conduct further research on how to best evaluate our UAC. We aim to evaluate a different use case with a larger participant-number. In future, we also want to further improve the connections in the mapping itself based on the experts' feedback. Thus we plan additional interviews with PPML experts to evaluate the whole framework. While we have now implemented the framework for a B2C use case, we are curious to investigate how the framework can be extended to B2B use cases. Therefore, our future main objective is an expert evaluated, advanced framework that is less time consuming.

8 Conclusion

In this paper, we have provided a process to systematically include user preferences for privacy preserving machine learning techniques (PPMLs) without asking for technical details. To do so we have suggested a framework that takes user preferences and developer knowledge as input to calculate a weighted score for each PPML. After providing the theoretical approach, we have applied the framework to a simplified PSI detection use case. While our previous work has focused on the mapping of User Acceptance Criteria (UAC) and PPML characteristics, this line of research focused on the collection of developer input and the sequential operation steps to calculate the weighted scores for each PPML technique.

Acknowledgements

This research received funding by the European Union (EU) - NextGenerationEU, and the German Federal Ministry of Education and Research under the BMBF project FIIPS-at-Home. The views expressed are those of the author(s) and do not necessarily reflect those of the EU or the EC.

References

- [1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138-52160, 2018.
- [2] Yoshimasa Akimoto, Kazuto Fukuchi, Youhei Akimoto, and Jun Sakuma. Privformer: Privacy-preserving transformer with mpc. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 392-410. IEEE, 2023.
- [3] Ahmad Al Badawi, Louie Hoang, Chan Fook Mun, Kim Laine, and Khin Mi Mi Aung. Privft: Private and fast text classification with homomorphic encryption. *IEEE Access*, 8:226544-226556, 2020.
- [4] Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, and Zumrut Muftuoglu. Privacy enabled financial text classification using differential privacy and federated learning. *arXiv:2110.01643*, 2021.
- [5] Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zumrut Muftuoglu, Sahib Singh, and Fatemehsadat Miresheghallah. Benchmarking differential privacy and federated learning for bert models. *arXiv:2106.13973*, 2021.
- [6] Majid Behzadian, S Khanmohammadi Otaghsara, Morteza Yazdani, and Joshua Ignatius. A state-of-the-art survey of topos applications. *Expert Systems with applications*, 39(17):13051-13069, 2012.
- [7] John H Benamati, Zafer D Ozdemir, and H Jeff Smith. An empirical test of an antecedents-privacy concerns- outcomes model. *JIS*, 43(5):583-600, 2017.
- [8] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv:1206.6389*, 2012.
- [9] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Trandic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *ECML PKDD, Proceedings, Part III 13*, pages 387-402, 2013.
- [10] Jonas Böhler and Florian Kerschbaum. Secure multi-party computation of differentially private heavy hitters. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2361-2377, 2021.
- [11] Amine Boulemtafes, Abdelouahid Derhab, and Yacine Challal. A review of privacy-preserving techniques for deep learning. *Neurocomputing*, 384:21-45, 2020.
- [12] Vanessa Bracamonte, Sebastian Pape, and Sascha Loebner. All apps do this": Comparing privacy concerns towards privacy tools and non-privacy tools for social media content. *PoPETs*, 3:57-78, 2022.
- [13] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245-317, 2021.
- [14] Di Cao, Shan Chang, Zhijian Lin, Guohua Liu, and Donghong Sun. Understanding distributed poisoning attack in federated learning. In *IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 233-239. IEEE, 2019.
- [15] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning attacks to local differential privacy protocols. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 947-964, 2021.
- [16] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.
- [17] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. Tem: High utility metric differential privacy on text. In *SIAM International Conference on Data Mining (SDM)*, pages 883-890. SIAM, 2023.
- [18] Ann Cavoukian. *Privacy by design*. 2009.

- [19] Konstantinos Chatzikokolakis, Miguel E. Andrués, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies: International Symposium*, pages 82-102, 2013.
- [20] Harsh Chaudhari, Matthew Jagielski, and Alina Oprea. Safenet: Mitigating data poisoning attacks on private machine learning. *arXiv:2205.09986*, 2022.
- [21] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv:1712.05526*, 2017.
- [22] European Commission, Content Directorate-General for Communications Networks, and Technology. Ethics guidelines for trustworthy AI. Publications Office, 2019.
- [23] Vincent Conitzer, Andrew Davenport, and Jayant Kalagnanam. Improved bounds for computing kemeny rankings. In *AAAI*, volume 6, pages 620-626, 2006.
- [24] Karolina Drobotowicz, Marjo Kauppinen, and Sari Kujala. Trustworthy ai services in the public sector: What are citizens saying about it? In *REFSQ 2021*, pages 99-115. Springer, 2021.
- [25] Sebastian Egger, Tobias Hossfeld, Raimund Schatz, and Markus Fiedler. Waiting times in quality of experience for web based services. In *2012 Fourth International Workshop on Quality of Multimedia Experience. IEEE*, 2012.
- [26] Jan-Erik Ekberg, Kari Kostinen, and Nadarajah Asokan. Trusted execution environments on mobile devices. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1497-1498, 2013.
- [27] European Parliament and Council of The European Union. Regulation (EU) 2016/679 General Data Protection Regulation (GDPR), 2016.
- [28] Weibei Fan, Jing He, Mengjiao Guo, Peng Li, Zhijie Han, and Ruchuan Wang. Privacy preserving classification on local differential privacy in data centers. *Journal of Parallel and Distributed Computing*, 135:70-82, 2020.
- [29] Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210-219. IEEE, 2019.
- [30] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC conference on computer and communications security*, pages 1322-1333, 2015.
- [31] J Marjan Hummel, John FP Bridges, and Maarten J IJzerman. Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. *The Patient-Patient-Centered Outcomes Research*, 7:129-140, 2014.
- [32] ISO/IEC 20889:2018-11. Privacy enhancing data de-identification terminology and classification of techniques. Standard, International Organization for Standardization, Geneva, CH, November 2018.
- [33] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1-2):1-210, 2021.
- [34] Fabian Karl and Ansgar Scherp. Transformers are short-text classifiers. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 103-122. Springer, 2023.
- [35] Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. Differential privacy in natural language processing: The story so far. *arXiv:2208.08140*, 2022.
- [36] Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, and Laurens van der Maaten. Crypten: Secure multi-party computation meets machine learning. *Advances in NeurIPS*, 34:4961-4973, 2021.
- [37] Lyudmila Kopeykina and Andrey Savchenko. Photo privacy detection based on text classification and face clustering. In *Proceedings of the VI International Conference Information Technology and Nanotechnology. Image Processing and Earth Remote Sensing (ITNT-IPERS 2020)*, volume 2665, pages 171-176, 2020.
- [38] Garam Lee, Minsoo Kim, Jai Hyun Park, Seung-won Hwang, and Jung Hee Cheon. Privacy-preserving text classification on bert embeddings with homomorphic encryption. *arXiv:2210.02574*, 2022.
- [39] Mu Li, Mingzheng Sun, Yijun Feng, and Xinyi Yang. Priber: Privacy-preserving entity resolution with secure multi-party computation. Available at SSRN 4453303.
- [40] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50-60, 2020.
- [41] Bill Yuchen Lin, Chaoyang He, Zihang Zeng, Hulin Wang, Yufen Huang, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. Fednlp: Benchmarking federated learning methods for natural language processing tasks. *arXiv:2104.08815*, 2021.
- [42] Ming Liu, Stella Ho, Mengqi Wang, Longxiang Gao, Yuan Jin, and He Zhang. Federated learning meets natural language processing: A survey. *arXiv:2107.12603*, 2021.

- [43] Xiaoyuan Liu, Hongwei Li, Guowen Xu, Zongqi Chen, Xiaoming Huang, and Rongxing Lu. Privacy-enhanced federated learning against poisoning adversaries. *Transactions on Information Forensics and Security*, pages 4574-4588, 2021.
- [44] Sascha Lvðbner, Boris Gogov, and Welderufael B Tesfay. Enhancing privacy in federated learning with local differential privacy for email classification. In *ESORICS 2022 International Workshops, DPM and CBT*, pages 3-18. Springer, 2023.
- [45] Sascha Lvðbner, Sebastian Pape, and Vanessa Bracamonte. User acceptance criteria for privacy preserving machine learning techniques. *ARES 2023*, 2023.
- [46] Sascha Lvðbner, Welderufael B. Tesfay, Vanessa Bracamonte, and Toru Nakamura. Systematizing the state of knowledge in detecting privacy sensitive information in unstructured texts using machine learning. *PST*, 2023.
- [47] Wen-jie Lu, Zhicong Huang, Zhen Gu, Jingyu Li, Jian Liu, Kui Ren, Cheng Hong, Tao Wei, and WenGuang Chen. Bumblebee: Secure two-party inference framework for large transformers. *Cryptology ePrint Archive*, 2023.
- [48] Zhuoran Ma, Jianfeng Ma, Yinbin Miao, Yingjiu Li, and Robert H Deng. Shieldfl: Mitigating model poisoning attacks in privacy-preserving federated learning. *Transactions on Information Forensics and Security*, pages 1639-1654, 2022.
- [49] Gianclaudio Malgieri. The concept of fairness in the gdpr: a linguistic and contextual interpretation. In *ACM FAccT 2020*, pages 154-166, 2020.
- [50] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information systems research*, 15(4):336-355, 2004.
- [51] Mohamed Merhi, Kate Hone, and Ali Tarhini. A cross-cultural study of the intention to use mobile banking between lebanese and british consumers: Extending utaut2 with security, privacy and trust. *Technology in Society*, 59, 2019.
- [52] Virraji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619-640, 2021.
- [53] Rakshit Naidu, Aman Priyanshu, Aadith Kumar, Sasikanth Kotti, Haofan Wang, and Fatemehsadat Mireshghallah. When differential privacy meets interpretability: A case study. *arXiv:2106.13203*, 2021.
- [54] Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1895-1904, 2022.
- [55] Mehta, D. B. (2023). Privacy-Preserving Machine Learning Architecture for Cross-Channel Advertising Optimization in a Large-Scale Social Media Platform. *International Journal of Communication Networks and Information Security (IJCNIS)*, 15(8), 23–32.
- [56] Robert Podschwadt and Daniel Takabi. Classification of encrypted word embeddings using recurrent neural networks. In *PrivateNLP@ WSDM*, pages 27-31, 2020.
- [57] B Thirumala Rao et al. A study on data storage security issues in cloud computing. *Procedia Computer Science*, 92:128-135, 2016.
- [58] Devin Reich, Ariel Todoki, Rafael Dowsley, Martine De Cock, et al. Privacy-preserving classification of personal text messages with secure multi-party computation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [59] Amanda Resende, Davis Railsback, Rafael Dowsley, Anderson CA Nascimento, and Diego F Aranha. Fast privacy-preserving text classification based on secure multiparty computation. *IEEE Trans. Inf. Forensics Secur.*, 17:428-442, 2022.
- [60] Marc Rothmann and Mario Pormann. A survey of domain-specific architectures for reinforcement learning. *IEEE Access*, 10:13753-13767, 2022.
- [61] Thomas L Saaty. Decision making-the analytic hierarchy and network processes (ahp/anh). *Journal of systems science and systems engineering*, 13:1-35, 2004.
- [62] Vasily A Sartakov, Daniel O'Keeffe, David Eysers, LluV's Vilanova, and Peter Pietzuch. Spons & shields: practical isolation for trusted execution. In *17th ACM International Conference on Virtual Execution Environments*, pages 186-200, 2021.
- [63] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3-18. IEEE, 2017.
- [64] Jiten Sidhpura, Parshwa Shah, Rudresh Veerkhare, and Anand Godbole. Fedspam: Privacy preserving sms spam prediction. In *International Conference on Neural Information Processing*, pages 52-63. Springer, 2022.
- [65] H Smith, Sandra Milberg, and Sandra Burke. Information privacy: Measuring individuals' concerns about organizational practices. *MIS quarterly*, pages 167-196, 1996.
- [66] Xiaoqiang Sun, Peng Zhang, Joseph K Liu, Jianping Yu, and Weixin Xie. Private machine learning classification based on fully homomorphic encryption. *IEEE Transactions on Emerging Topics in Computing*, 8(2):352-364, 2018.
- [67] Harry Chandra Tanuwidjaja, Rakyong Choi, Seunggeun Baek, and Kwangjo Kim. Privacy-preserving deep learning on machine learning as a service-a comprehensive survey. *IEEE Access*, 8:167425-167447, 2020.

- [68] Chandra Thapa, Jun Tang, Sharif Abuadbbba, Yansong Gao, Yifeng Zheng, Seyit Ahmet Váamtepe, Surya Nepal, and Mahathir Almashor. Fedemail: Performance measurement of privacy-friendly phishing detection enabled by federated learning. *ArXiv abs/2007.13300*, 2020.
- [69] Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. Ai robustness: a human-centered perspective on technological challenges and opportunities. *arXiv:2210.08906*, 2022.
- [70] Anh-Tu Tran, The-Dung Luong, Jessada Karnjana, and Van-Nam Huynh. An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation. *Neurocomputing*, 422, 2021.
- [71] Omkarprasad S Vaidya and Sushil Kumar. Analytic hierarchy process: An overview of applications. *European Journal of operational research*, 169(1):1-29, 2006.
- [72] Viswanath Venkatesh, James YL Thong, and Xin Xu. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, pages 157-178, 2012.
- [73] Michael Vierhauser and Manuel Wimmer. Towards integrating data-driven requirements engineering into the software development process: A vision paper. In *REFSQ 2020*, volume 12045, page 135. Springer Nature, 2020.
- [74] Hugo Villamizar, Amadeu Anderlin Neto, Marcos Kalinowski, Alessandro Garcia, and Daniel MV@ndez. An approach for reviewing security-related aspects in agile requirements specifications of web applications. In *RE 2019*, 2019.
- [75] Yongqin Wang, Rachit Rajat, and Murali Annavaram. Mpc-pipe: an efficient pipeline scheme for secure multi-party machine learning inference. *arXiv:2209.13643*, 2022.
- [76] Benjamin Weggenmann and Florian Kerschbaum. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *SIGIR 2018*, pages 305-314, 2018.
- [77] Chang Xu, Jun Wang, Francisco Guzmán, Benjamin Rubinstein, and Trevor Cohn. Mitigating data poisoning in text classification with differential privacy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.
- [78] Runhua Xu, Nathalie Baracaldo, and James Joshi. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv:2108.04417*, 2021.
- [79] Yabin Xu, Xuyang Meng, Yangyang Li, and Xiaowei Xu. Research on privacy disclosure detection method in social networks based on multi-dimensional deep learning. *Computers, Materials & Continua*, 62(1):137-155, 2020.
- [80] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv:2010.11947*, 2020.
- [81] Miaoyi Zeng, Shuaifu Lin, and Deborah Armstrong. Are all internet users' information privacy concerns (iuipc) created equal? *AIS TRR*, 6(1):3, 2020.
- [82]