Privacy-Aware Supervised Classification: An Informative Subspace based Multi-Objective Approach

Chandan Biswas [a,*], Debasis Ganguly[b], Partha Sarathi Mukherjee[c], Ujjwal Bhattacharya[a], Yufang Hou[d]

[a]*Indian Statistical Institute, Kolkata, India*
[b]*University of Glasgow, UK*
[c]*Tatras Data, Delhi, India*
[d]*IBM Research Europe, Dublin, Ireland*

Abstract

Sharing the raw or an abstract representation of a labelled dataset on cloud platforms can potentially expose sensitive information of the data to an adversary, e.g., in the case of an emotion classification task from text, an adversary-agnostic abstract representation of the text data may eventually lead an adversary to identify the demographics of the authors, such as their gender and age. In this paper, we propose a universal defense mechanism against such malicious attempts of stealing sensitive information from data shared on cloud platforms. More specifically, our proposed method employs an informative subspace based multi-objective approach to obtain a sensitive information aware encoding of the data representation. A number of experiments conducted on both standard text and image datasets demonstrate that our proposed approach is able to reduce the effectiveness of the adversarial

*Corresponding author
Email address: chandanbiswas08_r@isical.ac.in (Chandan Biswas)

task (i.e., in other words is able to better protect the sensitive information of the data) without significantly reducing the effectiveness of the primary task itself.

Keywords: Privacy preserving representation learning, Informative subspace, Multi-objective learning, Defence against information stealing adversarial attacks

---

1. Introduction

The era of data-driven learning is continuously witnessing increased computational requirements for training multi-layered complex neural networks for supervised machine learning (ML) through a layered approach of abstraction from the raw data, e.g., the work on contextual word vectors pre-trained on large collections of documents to capture the inherent language model in text [1], or that of training deep image networks to capture higher levels of visual features from images [2].

One standard solution to mitigate the intensive computational requirements of training data-driven models is to follow the standard 'software as a service' paradigm, in which the computations to train an ML model are provided as a service (MLaaS) by a powerful computing device (server), virtually accessible through a distributed computing environment (cloud) [3]. An MLaaS-based solution requires a user (client program) to upload an encoded form of the data, usually corresponding to an abstract representation of it, e.g. pre-trained vectors such as BERT [1] for text, or Inception-Net for images [2]), to the server. Although such an MLaaS based workflow allows provision for distributed data sharing and also reduces the computational overhead of the client workstations, a risk with an MLaaS architecture is that it can potentially lead to breaches in data security and privacy [4].

To illustrate the point on potential threats on data privacy, consider an *adversarial model* which is able to eavesdrop on the communication channel between a client and the server offering computation on encoded forms of data. Imagine a situation where an adversarial model is *pre-trained* on past data, which in terms of its domain and characteristics, is similar to the one that is transmitted to the server over a communication channel. In such a situation, this pre-trained adversarial model could use this submitted data as an input to predict a number of *sensitive* attribute values from this data [5].

As a concrete example of an adversarial attack on data privacy, consider that the encoded data sent from a client workstation to a computation server over a communication channel corresponds to that of movie reviews, and the *primary task* for which the computational resources of the server is sought, refers to the task of classifying a review into positive or negative, i.e. the primary task involves learning a mapping of the form $\theta : \mathbf{x} \mapsto y$, $\mathbf{x} \in \mathbb{R}^d, y \in \{0, 1\}$, where $\mathbf{x}$ represents an encoding of the data, e.g. a sequence encoding of the words comprising the review [6]. Imagine that each review contains additional *identity information attributes*, z, corresponding to sensitive information about the author, e.g. the age, gender etc. De-

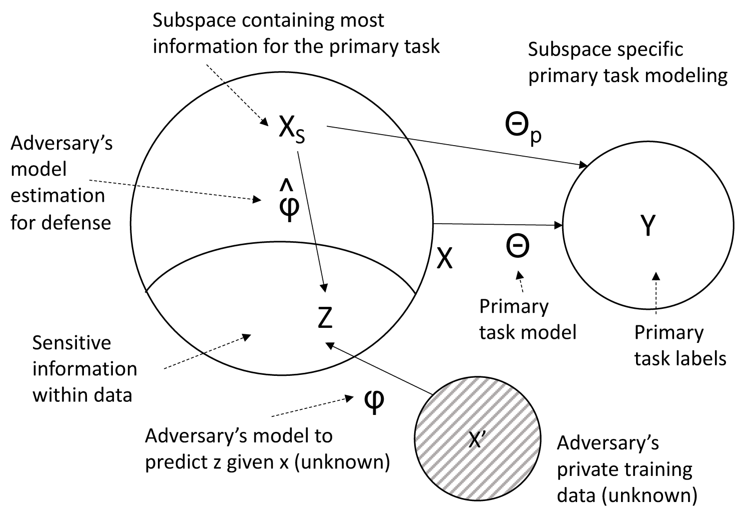


Figure 1: Schematic of our proposed defence mechanism that relies on identifying a candidate subspace, X_s, of the input space, on which the set of primary task labels, Y, is likely to exhibit a strong functional dependence. The remaining subspace, $X - X_s$, is then useful to estimate a likely functional dependence with the sensitive information, $\hat{\phi}$, an inversion on which is then used to defend against an adversarial model, ϕ.

spite not being a part of the encoding, the adversary can potentially feed the encoded data as input into an adversarial network, that has already been trained on pairs of movie reviews encoding and the attribute values (e.g. gender), $(\mathbf{x}', z)$, to learn an association between the two of the form $\phi : \mathbf{x}' \mapsto z$, $\mathbf{x}' \in \mathbb{R}^d, z \in \{0, 1\}$. The parameters of the trained network, ϕ, may then accurately predict the demographics of the current encoded data $\mathbf{x}$, i.e., the closer $\mathbf{x}$ is to $\mathbf{x}'$ the higher is the associated risk of leaking the attribute value information [7].

A standard approach to prevent an attacker stealing the sensitive information from data is to make the encoding process itself aware of the intentions of an adversary, which usually involves first formulating the adversarial model, $\phi : \mathbf{x} \mapsto z$ as a secondary task, and then applying a multi-objective based encoding transformation of the data, where the first objective corresponds to the primary task and the subsequent ones correspond to one or more secondary tasks, each such secondary task representing an adversarial objective [5]. The learning objective, in this case, seeks to minimize the potential degradation of the primary task effectiveness due to the noise which is required to be incorporated within the data as a defence against adversarial attacks.

Our Contributions. We now enlist our contributions in this paper. First, contrary to a standard approach of data-driven encoding that uses uniform weights for the abstract features, we hypothesize that the defence mechanism of a multi-objective based approach can potentially be improved by a weighted distribution over features. Specifically, this involves leveraging information from *candidate subspaces*, $\mathbf{x_s} \in \mathbb{R}^k, (k < d)$ of the input data that are *strongly correlated with the primary category labels* in the form $\theta_p : \mathbf{x_s} \mapsto y$. The *residual subspace* is thus likely to be functionally associated to the latent attribute values of the data, or in other words, to the secondary (adversarial) task categories $\hat{\phi} : \mathbf{x_s}' \mapsto z, \mathbf{x_s}' \in \mathbb{R}^{d-k}$, which in turn approximately models the function $\phi : \mathbf{x} \mapsto z$. We argue that this way of modeling the adversarial information yields a more robust encoding mechanism that is likely to be more resilient to security threats and our experiments confirm this hypothesis.

Second, in contrast to most existing approaches which conduct experiments mostly on text data with annotated metadata information (such as the demographic attributes, e.g., age and gender annotated as a part of the TrustPilot dataset [5]), we report empirical results on both images and text. For images, we test our method both on implicit and explicit demographic attributes. As implicit attributes, we use stylistic attributes, such as the slant or ligatures in handwriting, that could potentially reveal the age of a person. As explicit attributes, we test if the metadata information of age and gender associated with a set of lesion images can potentially be revealed to information stealing attacks.

2. Related Work

Adversarial Learning. An adversarial attack broadly refers to the methods of generating samples (often called adversarial examples) that are indistinguishable from samples drawn from the true data distribution with an objective to 'fool' a classifier [8]. These attacks typically use first order gradient information, such as FGSM [8], I-FGSM [9], MI-FGSM [10], Ada-FGSM [11] etc. Successful demonstrations of black-box adversarial perturbations attacks leading to degrading the effectiveness of classifiers were demonstrated in [12] and [13]. Defence mechanisms against such adversarial attacks include those of using regularized FGSM [14], and defensive distillation [15].

Different from adversarial learning, we rather employ a multi-objective encoding, the purpose of which is to ensure that it potentially would be difficult for an adversary to use a pre-trained system (on similar data) to effectively predict the values of sensitive attributes (e.g., age, gender etc.)

from the encoded data.

Differential Privacy and Privacy-preserving Data Encoding. The objective of differential privacy is somewhat similar to that of privacy-preserving encoding. However, differential privacy does not involve encoding the raw data as vectors; instead, it obfuscates parts of relational data so as to mitigate individual data leakage [16]. Various de-identification or anonymizing technologies have been proposed to protect data privacy, which often involve adding noise or masking sensitive information in the released dataset [17]. The concept of additive noise in differential privacy for relational databases also finds applications in Bayesian risk minimization in general [18], or in Bayesian linear regression [19] in particular. Privacy preserved data encoding finds applications in encoding raw data for both unsupervised [20] and supervised learning tasks [21]. For text data, privacy-preserving based encoding is particularly crucial because the inherent characteristics of natural language (e.g., writing style or word usage patterns) often reveal information about the authors, which can be used by adversaries to reveal such sensitive information. As examples, the authors of [22] used online behavior, stylistic choices and language models to predict the age group of blog authors, while those of [23] used Twitter content to predict the occupational class.

A number of recent studies has proposed the dual objective of privacy preservation (minimizing leakage of sensitive information) and model preservation (maximizing the performance of an algorithm on the encoded data), e.g., applying a 'multi-detasking' model to train an adversarial classifier simultaneously with the primary downstream text classifier, where during training, the primary classifier updates its parameters to confuse the attacker model [5]. The study reported in [24] developed a distributed framework for privacy preserving multi-task learning protocol by applying encryption mechanisms. The authors of [4] explored an adversarial learning approach that learns unbiased representations of text with respect to specific sensitive attributes. Somewhat different from the findings of [5], the authors of [25] showed that despite adversarial training methods being generally effective in reducing the amount of implicit sensitive information, in some cases, however, a substantial amount of sensitive information still persists and can be extracted from the encoded representations.

Although our proposed method falls into the general class of multi-objective approaches, such as those of [5] and [26], our proposed method is more general in the sense that we leverage the candidate subspaces that are most informative of the primary task. Since parts of these subspaces are less likely to be comprised of the sensitive information in data, our method seeks to address some of the concerns pointed out in [25], i.e. removal of sensitive attributes (e.g. demographics) from data instances can still lead to an adversary predicting this missing information. Our subspace-based approach is explicitly directed towards mitigating this problem in the sense that the privacy-aware encoding process puts more emphasis only on those components of the data that are more useful for the primary task, while suppressing the residual space that contains most of the information on the sensitive attributes.

Feature Importance for Explanations. Standard approaches of model-agnostic instance-wise explanations for classification include those of employing linear regression to learn a simplified decision boundary by sampling points around a data instance [27], applying a Gumbel distribution to estimate instance-wise feature importance [28] etc. The authors of [29] reiterate the importance of feature selection for supervised learning tasks, whereas those of [30] and [31] explore feature selection for the case of unsupervised learning.

In the context of our work, we use the idea of exploring informative candidate subspaces with a parameterized approach, as first proposed in [28]. An explicit use of feature importance also provides an interpretable way of preserving data privacy.

3. A General Framework for Privacy-Aware Encoding

In this section, we formally describe a general framework for defence against adversarial threats using a multi-task learning based workflow. We present a general approach to the problem in the sense that the overall framework allows provision to incorporate more than one adversarial task, each corresponding to a particular attribute of the data.

3.1. Privacy-Agnostic Encoding

Using the notations introduced Section 1, the predictive model for the primary task, generally speaking, can be *learned* with a set of linear transformation functions (realized with a multi-layer perceptron) of the form

$$P(y = i|\mathbf{w};\theta, \theta_p) = \sigma(\theta_p \cdot \mathbf{x})_i = \frac{\exp(\theta_{p_i} \cdot \theta \cdot \mathbf{w})}{\sum_{j=1}^{c} \exp(\theta_{p_j} \cdot \theta \cdot \mathbf{w})},$$
$$\mathbf{x} = \theta \cdot \mathbf{w},\ \mathbf{x} \in \mathbb{R}^s, \mathbf{w} \in \mathbb{R}^d,\ y \in \mathbb{Z}_c, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ denotes a d-dimensional vector representation (encoding) of the input data, $y \in \mathbb{Z}_c$ denotes a class label (one of c possible values) corresponding to the classification task, $\theta \in \mathbb{R}^{s\times d}$ denotes a matrix of parameters (a latent layer of a neural network), and $\theta_p \in \mathbb{R}^{c\times s}$ denotes a matrix of parameters specifically corresponding to the classification task ($\theta_{p_i} \in \mathbb{R}^s$ is the parameter vector for the i-th class). As a simplification, we do not explicitly include the bias parameter as a part of the softmax equations. Since the encoding process of Equation 1 does not explicitly take account an adversarial threat against a subset of data attributes, the encoding $\mathbf{x} \in \mathbb{R}^s$ is privacy-agnostic.

3.2. Privacy-Aware Encoding

An encoding space different from Equation 1 that explicitly addresses a set of sensitive attributes has been shown to be effective in defence against adversarial models [5]. However, the work in [5] addresses the defence mechanism for a single attribute only. Instead, we present a more general setup involving more than one attribute.

In the context of our work, the attributes manifest themselves as an implicit part of the data, or otherwise, it is straight-forward to remove the attributes before encoding the data [25]. In particular, we assume that the encoding of an input data instance, $\mathbf{w}$, is a function of both the raw data itself, (say w) and its latent characteristics (sensitive attributes). We represent a pair comprising an input data instance and a set of M sensitive attributes (assuming categorical values) associated with it as $(w, \{z_1, \ldots, z_M\})$, where $z_m \in \mathbb{Z}_{s_m}$, i.e. there are a total of s_j number of possible values for the j^{th} attribute.

A multi-objective transformation then uses the pairs, $(w, \{z_1, \ldots, z_M\})$, to encode the privacy-agnostic representation $\mathbf{w} \in \mathbb{R}^d$ as learnable parameters, $\mathbf{x} \in \mathbb{R}^s$, with the combined objective

$$\begin{aligned} P(y = i, z_1, \ldots, z_M | \mathbf{w}; \theta, \theta_p, \phi^1, \ldots, \phi^M) = \\ (1 - \sum_{m=1}^{M} \gamma_m) \sigma(\theta_p \cdot \mathbf{x})_i - \sum_{m=1}^{M} \gamma_m \sigma(\phi^m \cdot \mathbf{x})_{z_m}, \end{aligned} \tag{2}$$

where $\mathbf{x} = \theta \cdot \mathbf{w}$, $\mathbf{x} \in \mathbb{R}^s$, and $\mathbf{w} \in \mathbb{R}^d$, and similar to Equation 1, $\sigma(.)_i$ is an abbreviation for the softmax function with respect to the i-th class. The multi-objective loss of Equation 2 can be realized with a feed-forward network comprising a shared layer (parameter matrix $\theta \in \mathbb{R}^{s \times d}$) and the task specific layers. Separate layers, one for each adversarial task ($\phi^m \in \mathbb{R}^{s_m \times s}$), in addition to the primary task itself ($\theta_p \in \mathbb{R}^{c \times s}$), are all connected to the shared layer. Note that the parameters corresponding to $\mathbf{w}$'s in Equation 2 are obtained from pre-trained representations and hence are not learnable.

To illustrate Equation 2 with an example, consider a text classification problem, where each document is associated with the demographic attributes - age (z_1) and gender (z_2) of author. In such a situation, the value of M in Equation 2 would be 2. Continuing with the example, if age is discretized into 3 categories, e.g., 'young', 'middle-aged' and 'senior' then $s_1 = 3$.

In a generalized setting, the multi-objective loss function of Equation 2 models a relative trade-off between the effectiveness of the primary task and the desired lack of effectiveness of the adversarial ones (notice the negative factor in the linear combination corresponding to the adversarial tasks). A low value of each linear combination parameter, $\gamma_m \in [0,1] : (\sum_m \gamma_m < 1)$, associates a small importance to the necessity of defending against an information stealing attack against the m-th attribute. Notice that setting $\gamma_m = 0$ degenerates Equation 2 to the privacy-agnostic encoding of Equation 1.

4. An Information Theoretic Perspective

In this section, we describe how to extend the general multi-task based privacy preserving approach from an information theoretic perspective. As per the motivation behind the schematic depiction of Figure 1, we now formally describe how to leverage information from the importance of features (components of the encoded vector representation of a data instance) to help the process of learning a better encoding for privacy preservation.

4.1. Subspace Encoding

A limitation of Equation 2 is that the parameters of the shared layer and the primary-task specific layer (i.e. θ and θ_p respectively) are trained with respect to the entire feature space of the encoded vector $\mathbf{w}$, whereas it is more likely to be the case that a part of this feature space correlates strongly with the primary task. The key idea in our proposed method is to substitute the encoding $\mathbf{w}$ of Equation 2 with a subset of features that are most likely to be informative for the primary task. This has a two-fold advantage.

First, a subspace of the most informative features for the primary task is likely to lead to a down-weighting of the residual subspace potentially constituting information responsible for determining the values of the sensitive attributes of the data. In other words, this is likely to degrade the effectiveness of the secondary tasks thus providing a potentially improved defence mechanism.

Second, since the subspace-based encoding approach puts more emphasis on parts of the data that are potentially responsible for determining the primary task output, it is also likely to lead to improving the effectiveness of the primary task itself.

4.2. Parameterized Subspace Selection with Gumbel Distribution

The authors of [28] computed the importance of features by measuring the mutual information between the primary task labels and an arbitrary feature subspace $\mathbf{w_s} \in \mathbb{R}^k, (k < d)$. The total number of possible subspaces,

$\binom{d}{k}$, is exponential for relatively large values of k. Hence finding an optimal subspace representing the largest amount of information for data driven models is a challenging problem. A solution, proposed in [28, 32], is to use a parameterized version of a subspace (specifically obtained with a Gumbel distribution) that allows a gradient descent based optimization of its parameters. The objective is seek an optimum state of maximum informativeness of the subspace with respect to a set of labels. Before describing how this is applied in the context of our problem, we present a brief overview of the Gumbel based learning of subspaces, mostly following the exposition of [28].

A Gumbel distribution, $G(0,1)$, is a distribution of random variables of the form $G_i = -\log(-\log u_i), u_i \sim \mathcal{U}(0,1), \mathcal{U}$ being the uniform distribution. The Gumbel softmax probability distribution uses a concrete distribution, which is a continuous differentiable approximation of a categorical random variable. The *Gumbel softmax* is a modification of the softmax function involving random variables sampled from the Gumbel distribution, one each for each component of the softmax. In the context of our problem, we use the Gumbel softmax distribution to estimate the importance of each component of the encoding vector, $\mathbf{w} \in \mathbb{R}^d$. Formally speaking,

$$C = \{C_i : C_i = \frac{\exp((\log w_i + G_i)/\rho)}{\sum_{j=1}^{d} \exp((\log w_j + G_j)/\rho)},\ i = 1, \ldots, d\}, \tag{3}$$

where ρ is a *temperature* parameter, higher values of which makes the distribution close to uniform (for our experiments, we set $\rho = 0.1$ as per [28]).

To select k features from a set of available d features, one needs to independently sample k times from the Gumbel softmax distribution resulting in a total of k random vectors $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$, where the j^{th} vector $\mathbf{c}_j$ is sampled from Gumbel softmax, i.e., $\mathbf{c}_j \sim C$. Let $\Lambda_k \in \mathbb{R}^{d \times k}$ be the matrix constituted from the k random vectors, $\mathbf{c}_j$, thus sampled. A row-wise maximum of the matrix, Λ_k then yields an approximation of a k-hot random vector $\lambda_k \in \mathbb{R}^d$. The highest k elements of λ_k (corresponding to the most important features) are retained while the rest $(d-k)$ are set to 0. Thus λ_k is a vector with k non-zero elements (*soft k-hot*) determining the choice of a k-dimensional subspace.

4.3. Feature Subspace with Multi-Objective

In the context of our problem (see Equation 2), data is represented as vectors in d dimensions, i.e. $\mathbf{w} \in \mathbb{R}^d$, out of which we intend to select a subspace $\mathbf{w_s} \in \mathbb{R}^k$ comprised of the most informative features. After selecting a random vector with k non-zero elements, λ_k, we now model its interaction with the primary classification task as

$$\begin{aligned} P(y = i, z_1, \ldots, z_M | \mathbf{w}; \theta, \theta_p, \phi^1, \ldots, \phi^M) = & \\ (1 - \sum_{m=1}^{M} \gamma_m) \sigma(\theta_p \cdot \mathbf{x})_i - \sum_{m=1}^{M} \gamma_m \sigma(\phi^m \cdot \mathbf{x})_{z_m}, & \end{aligned} \tag{4}$$

where $\mathbf{x} = \theta \cdot (\mathbf{w} \odot \lambda_k)$, $\mathbf{x} \in \mathbb{R}^s$ and $\mathbf{w} \in \mathbb{R}^d$. Equation 4 is a more constrained form of Equation 2. This is because instead of considering an arbitrary s-dimensional transformation from $\mathbf{w}$ (privacy-agnostic encoding) to $\mathbf{x}$

(privacy-aware encoding) of Equation 2, we specifically select an informative subspace, denoted by, say $\mathbf{w_s} = \mathbf{w} \odot \lambda_k$. This is obtained by an element-wise multiplication of the input encoding with a soft k-hot vector obtained from the Gumbel softmax distribution.

As a next step, the informative subpace is used to learn the privacy-aware encoded representation[1]. In our experiments, instead of specifying the value of k directly, we control it with a fraction, $\tau \in [0, 1]$ of the input data dimension, i.e., $k = \lfloor \tau d \rfloor$.

5. Experimental Setup

5.1. Experiment Workflow

A laboratory based setup is devoid of the presence of a true adversary (e.g. as shown in the schematic of Figure 1). In such a situation, the adversary would have access to a pre-trained model which is trained to predict the sensitive attributes from input data instances. An adversarial model is likely to be more harmful if it has been trained on data instances that resemble the ones (i.e. similar in terms of encoded vector representations) to the ones that are sent over from the client to the MLaaS. To mimic this situation as closely as possible in a laboratory setup, we set up our experiments as shown in Figure 2.

For each labeled dataset, each data instance is annotated with additional

[1]A prototype of the implementation is available for research purposes at `https://github.com/chandanbiswas08/l2x-mt`

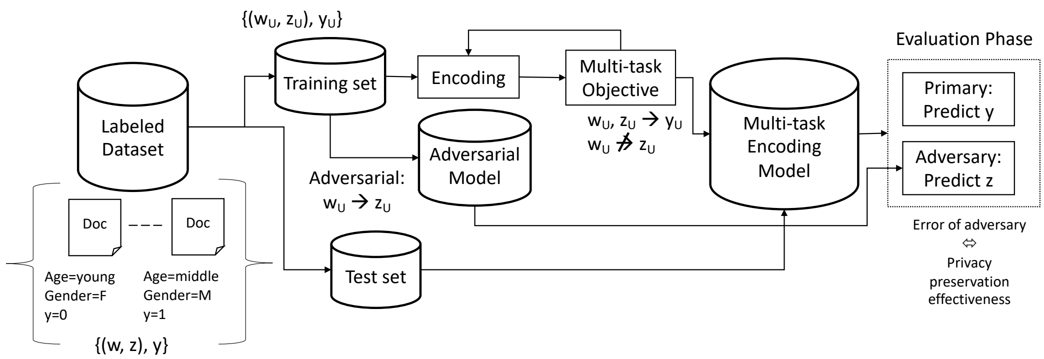


Figure 2: Schematics of the common setup for the evaluation workflow. Both the privacy-aware encoding and the adversarial model (one for each attribute) is trained on the train-split of the data. During evaluation phase, the *privacy-preserved* encoded vectors for the test-split are fed into the adversarial model to predict values of the attributes. The prediction error of this pseudo-adversarial setup indicates the effectiveness of privacy preservation.

attribute value pairs. With this we train a logistic regression model on the train-split of the data to simulate an adversarial attack of predicting these additional attribute values from the data (a separate adversarial model is trained for each attribute type, shown as a single model in Figure 2 to avoid clutter).

In general, corresponding to M different attribute types (see Equations 2 and 4), we evaluate the effectiveness of the adversarial task as an *inverse effectiveness* measure for a particular defence method used in our experiments. The experiment workflow ensures that the encoding process of a defence mechanism is oblivious of the category values (e.g., values of age and gender) of the test-split.

Table 1: Summary of the dataset used in our experiments.

Dataset	#Instances Train	#Instances Test	Primary task Classify	Primary task #Classes	Adversarial Tasks Type	Adversarial Tasks Attribute	Adversarial Tasks Categories
Morpho-MNIST (M-MNIST)	120K	40K	Digits	10	Synthetic	Slant Fracture	{left, neutral, right} {yes, no}
Skin Cancer MNIST (HAM10K)	8500	1500	Diseases	7	Real	Age Gender	{$\leq$ 30, 31-60, $>$ 60} {male, female}
Trustpilot (US English)	23K	4K	Sentiment	2	Real	Age Gender	{$\leq$ 35, $>$ 35} {male, female}

5.2. Dataset

A dataset suitable for the purpose of our experiments needs to be annotated with additional attribute values corresponding to the sensitive information, the prediction of which during the adversarial workflow branch (see Figure 2) could then be set up as information leakage. To test the effectiveness of our proposed subspace based privacy preservation approach on different modalities of data, we experiment with both text and image datasets. The details of each dataset follows next (also summarized in Table 1).

Morpho-MNIST (M-MNIST). The primary task of the original MNIST dataset involves detecting the class of a digit (a gray-scale image with 28×28 pixels) out of the 10 possibilities (one of 0 to 9). As a part of latent information that can potentially be leaked from an encoding of a hand-written image (e.g. a 2d convolution with maxpooling), we first consider the *slant* of a hand-written digit, which can be considered to be correlated with personality traits [33]. To setup the dataset, each *slant label*, z_1 (in our notation), is obtained by applying a threshold on the horizontal shear, α. The value of the shear, α, in turn is computed as a function of second order moments of the gray-scale values, x_{ij} [34]. Formally,

$$z_1 = \begin{cases} 0 & \alpha \leq -0.3\,(\text{left}) \\ 1 & -0.3 < \alpha < 0.3\,(\text{neutral}) \\ 2 & \alpha \geq 0.3.(\text{right}) \end{cases} \tag{5}$$

In addition to the slant, the second attribute that we address in our experiments is whether the image of a hand-written digit is *fractured*, i.e., a lack of continuity is exhibited in the strokes. The value of this attribute, if revealed in a real-life situation, could indicate the age of an OCR-ed document to an adversary.

For our experiments with the fracture attribute, we use an existing dataset, namely the 'Morpho-MNIST', where morphological erosion is applied to synthetically generate fractured images [34]. Addition of the synthetically generated fractured images, one for each image in the original MNIST, resulted in doubling the number of images for this dataset. The information on whether an image is fractured is not available to an adversary, nor does the adversary is allowed to compute the slant labels using Equation 5.

Skin Cancer MNIST (HAM10K). Contrary to using synthetically generated attribute values for the adversarial task, the 'Skin Cancer MNIST' (or HAM10K) dataset [35] allows us to setup the adversarial tasks with two explicitly annotated attributes. The primary task in this dataset involves identifying one out of 7 possible skin diseases, e.g., Bowen's disease, basal

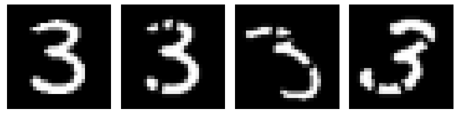


Figure 3: Left to right: No slant and fractures, followed by fractures with neutral, left and right slants.

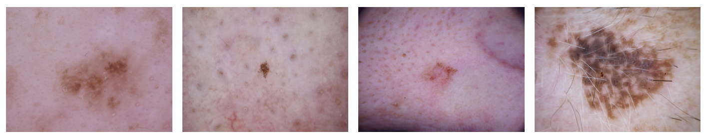


Figure 4: Left to right: Lesion images of a young female, mid-aged male, old female and an old male.

cell carcinoma etc., from images of lesions. The objective in this case is to encode the data in such a way that it does not reveal the age or gender of a person without substantially degrading the effectiveness of the primary task. Some sample images from the two image datasets are shown in Figures 3 and 4.

TrustPilot Dataset. For the text modality, we use the TrustPilot reviews (the US English subset). The primary task on this dataset involves identifying sentiment (positive or negative) of a review [36]. This dataset, comprised of over $27K$ reviews with sentiment score ranging between 1 and 5, has annotated values for both age and gender. Since the number of reviews with scores 2 and 3 is substantially small, we binarize the sentiment class labels by thresholding with a value of 3, i.e. scores from 1-3 are mapped to class 0 and the rest to 1. Following the previous experiment setup of [5] and [4], we binarize the attribute 'age' as young (age $\leq$ 35) and its complement (representing the category 'not young').

5.3. Baselines

As baselines, we compare the following approaches. First, we apply a privacy agnostic logistic regression based approach (see Equation 1), which we denote as **LR**. Our next baseline, denoted as **MT**, is the multi-tasking based approach from existing literature [5], which we presented in this paper as Equation 2. To explore if subspace based information usage, which forms a part of our proposed method, is indeed effective, we conduct experiments with two ablation baselines. The first of these baselines (applicable for text) involves the following. After computing the term feature weights with a simple term importance statistics (specifically tf-idf), for each sentence we retain only a fraction, $\tau \in [0, 1]$, of the terms with the highest weights. The rationale of this baseline, denoted as **LR-TFIDF**, is to see if removing a subset of features, not correlated to the primary task alone, can prevent information leakage of secondary attributes.

The second ablation baseline is a degenerate case of Equation 4, where we set $\gamma_m = 0$ for each adversarial task. This means that the k-dimensional encoding of the data, being agnostic of the adversarial tasks, only takes into account the informative subspace of the primary task. Unlike LR-TFIDF, this baseline method, denoted as **L2X** in our experiments, is applied to both text and images.

5.4. Evaluation Metrics and Parameters

As an evaluation metric, we employ a combination of the primary task accuracy (higher the better) and the inverse accuracy of the secondary tasks (lower the better). A high value of the combined metric reflects a better defence against information leakage without a substantial drop in primary task effectiveness. For combination, we specifically use the harmonic mean between the inverse of the aggregated accuracy values of the secondary tasks and the accuracy of the primary task, i.e.,

$$F_S = \frac{2A_P(1-A_S)}{(1-A_S)+A_P}, \tag{6}$$

where each A_S is the harmonic mean over the accuracy of each adversarial task, A_{S_i}.

The hyper-parameters tuned for each method were: a) τ, which controls the number of features retained (for the LR-TFIDF baseline, this refers to the fraction of the terms retained with the highest tf-idf scores), and b) (γ_1, γ_2), which controls the relative importance of the two adversarial tasks (Equation 4). In particular, the range of these hyper-parameters in our experiments were: $[0.2, 0.8]$ for τ, and $[0.1, 0.4]$ for γ_1 and γ_2, in steps of 0.2 and 0.1 respectively.

5.5. Results

Summary. Table 2 summarizes the optimal results of the different privacy preservation learning methods. The optimal result for each method was

Table 2: Privacy preservation results summary on different datasets. Parameter combinations that are not applicable for a method are shown as filled up gray cells. e.g. the parameter γ_1 for LR.

		Hyper-parameters			Accuracy			Combined Measures		
Dataset	Method	τ	γ_1	γ_2	A_P	A_{S_1}	A_{S_2}	F_{S_1}	F_{S_2}	F_S
TrPilot	LR				0.8674	0.7292	0.7168	0.4127	0.4270	0.4200
	LR-TFIDF	0.2			0.8194	0.7113	0.6928	0.4270	0.4469	0.4371
	MT		0.4	0.4	0.8694	0.6849	0.6920	0.4626	0.4549	0.4587
	L2X	0.2			**0.8726**	0.6804	0.6546	0.4678	0.4949	0.4818
	L2X-MT	0.6	0.1	0.1	0.8711	**0.6564**	**0.6465**	**0.4928**	**0.5029**	**0.4979**
M-MNIST	LR				0.9840	0.8956	0.6992	0.1888	0.4608	0.3525
	MT		0.2	0.2	**0.9851**	0.8647	0.6735	0.2379	0.4904	0.3896
	L2X	0.4			0.9593	0.5435	0.5764	0.6186	0.5877	0.6038
	L2X-MT	0.4	0.4	0.1	0.9596	**0.5291**	**0.5420**	**0.6318**	**0.6201**	**0.6260**
HAM10K	LR				0.6995	0.5757	0.6256	0.5282	0.4877	0.5093
	MT		0.3	0.2	**0.7072**	0.5749	0.6249	0.5310	0.4902	0.5119
	L2X	0.2			0.6861	0.5384	0.6045	0.5519	0.5018	0.5290
	L2X-MT	0.6	0.4	0.4	0.6861	**0.5376**	**0.6017**	**0.5525**	**0.5040**	**0.5303**

obtained by individually tuning its hyper-parameters.

We observe that although LR, being a privacy agnostic approach, results in high effectiveness for the primary task classification, it also yields high values for the adversarial tasks. This indicates a substantial information leakage with the LR method. Multi-tasking based encoding (MT) helps improve results, specially for text, as also noted in [5].

Subspace encoding alone (L2X) is also able to decrease the accuracy values for the adversarial tasks (i.e. improve privacy preservation), which also means that a combination of MT and L2X should also improve results. This is precisely what is demonstrated by the results of our method (L2X-MT), which yields the best results for each dataset.

Parameter Sensitivity. We also investigate the effects of varying τ (subspace selection), and the relative importance of the adversarial task (γ_m) parameters (Equations 2 and 4) on the overall effectiveness of privacy-preservation

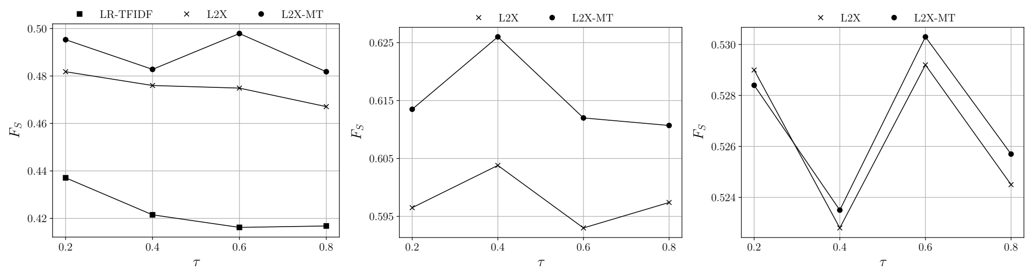


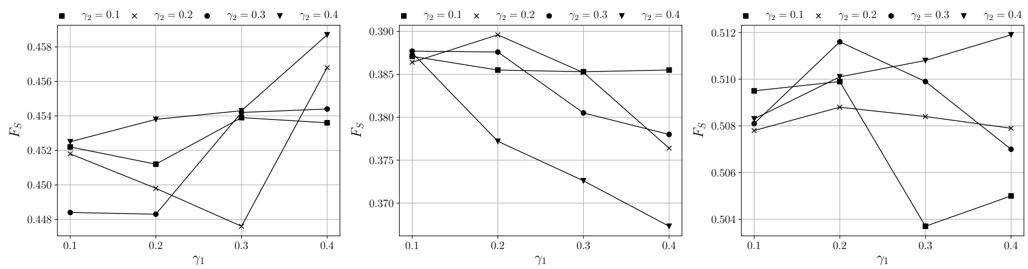
Figure 5: Sensitivity of the privacy-aware learning approaches with respect to relative subspace dimensionality τ; Left: TrustPilot, Middle: M-MNIST, Right: HAM10K.

learning of the corresponding primary tasks. Figure 5 shows that L2X-MT outperforms the baselines consistently for a range of different subspace dimensions. Figure 6 shows the relative comparisons between the two multi-tasking approaches - MT and L2X-MT. It can be seen that for a range of different relative importance of the two adversarial tasks (e.g. age/gender detection for Trustpilot and HAM10K, and slant/fracture detection for M-MNIST), leveraging information from informative subspaces helps improve the overall balance between primary task effectiveness and prevention of information leakage.

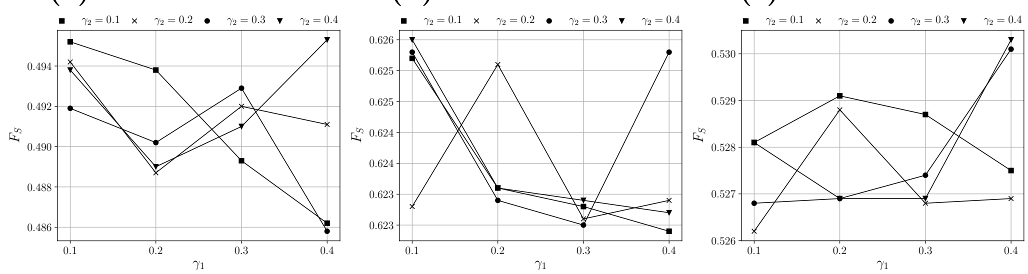
In summary, our experiments revealed the following two key observations.

1. Learning on data encoded by our method yields comparable results with that obtained on data in its original form, i.e. *our proposed encoding does not lead to a significant decrease in the effectiveness of a classification model*.

2. Data encoded by our method *significantly reduces the effectiveness of an adversarial classification model* which seeks to predict sensitive attributes from the data. It is also shown that the use of the *informative*



(a) TrustPilot: MT **(b)** M-MNIST: MT **(c)** HAM10K: MT



(d) TrustPilot: L2X-MT **(e)** M-MNIST: L2X-MT **(f)** HAM10K: L2X-MT

Figure 6: Sensitivity of MT, L2X-MT with variations in relative importance of two adversarial tasks.

subspace helps to improve the defence mechanism, i.e., it further reduces the effectiveness of the adversarial classification model.

6. Conclusions and Future Work

We proposed a generic method of privacy-preserving supervised learning, which is potentially beneficial for distributing an encoding of the input data over a cloud environment with the end-goal of eventually learning a predictive model (primary task) on the data. Our generic methodology combines the advantages of two main hypotheses - that of (a) using a *multi-task objective* that in addition to learning the primary task also learns the complementary (inverse) characteristics of an adversarial model as a defence mechanism against information stealing attacks; and (b) using a *residual subspace* of the data to further improve the defence mechanism.

Our experiments on image and textual data demonstrated that our proposed method, which jointly learns a multi-objective encoding over informative subspaces (with respect to the primary task), outperforms a separate application of each.

In future, we would like to explore how may it be possible to obtain a privacy-preservation encoding of the input data in those cases where the sensitive attributes are latent rather than being manifested as explicitly annotated identifiable attributes (i.e., to address the situation when the attribute value annotations are not available in the training set). Unsupervised analysis of the input space coupled with a semi-supervised encoding approach can potentially be useful to tackle such a situation.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[2] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.

[3] M. Ribeiro, K. Grolinger, M. A. Capretz, Mlaas: Machine learning as a service, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), IEEE, 2015, pp. 896–902.

[4] Y. Li, T. Baldwin, T. Cohn, Towards robust and privacy-preserving text representations, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 25–30.

[5] M. Coavoux, S. Narayan, S. B. Cohen, Privacy-preserving neural representations of text, in: Proc. of EMNLP '18, 2018, pp. 1–10.

[6] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proc. of ICML'14, 2014, pp. II–1188–II–1196.

[7] B. Weggenmann, F. Kerschbaum, Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining, in: ACM SIGIR '18, 2018, pp. 305–314.

[8] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Y. Bengio, Y. LeCun (Eds.), ICLR '15, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[9] A. Kurakin, I. J. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: ICLR '17, Toulon, France, April 24-26, 2017, Workshop Track Proceedings, 2017.

[10] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proc. of CVPR '18, 2018, pp. 9185–9193.

[11] Y. Shi, Y. Han, Q. Zhang, X. Kuang, Adaptive iterative attack towards explainable adversarial robustness, Pattern Recognition (2020) 107309.

[12] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 506519.

[13] D. Li, J. Zhang, K. Huang, Universal adversarial perturbations against object detection, Pattern Recognition 110 (2021) 107584.

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: Y. Bengio, Y. LeCun (Eds.), Proc. of ICLR'14, 2014.

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: 2016 IEEE Symposium on Security and Privacy (SP), IEEE, 2016, pp. 582–597.

[16] C. Dwork, Differential privacy, in: 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Vol.

4052 of Lecture Notes in Computer Science, Springer Verlag, 2006, pp. 1–12.

[17] R. Wang, B. C. Fung, Y. Zhu, Q. Peng, Differentially private data publishing for arbitrarily partitioned data, Information Sciences 553 (2021) 247–265.

[18] C. Dimitrakakis, B. Nelson, Z. Zhang, A. Mitrokotsa, B. I. Rubinstein, Differential privacy for bayesian inference through posterior sampling, JMLR 18 (1) (2017) 343–381.

[19] G. Bernstein, D. R. Sheldon, Differentially private bayesian linear regression, in: Proc. of NIPS '19, 2019, pp. 525–535.

[20] C. Biswas, D. Ganguly, D. Roy, U. Bhattacharya, Privacy preserving approximate k-means clustering, in: Proc. of CIKM '19, 2019, pp. 1321–1330.

[21] Y. Jinfeng, W. Jun, J. Rong, Privacy and regression model preserved learning., in: Proc. of AAAI '14, 2014, pp. 1341–1347.

[22] S. Rosenthal, K. McKeown, Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 763–772.

[23] D. Preoţiuc-Pietro, V. Lampos, N. Aletras, An analysis of the user occupational class through twitter content, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 1754–1764.

[24] K. Liu, N. Uplavikar, W. Jiang, Y. Fu, Privacy-preserving multi-task learning, in: Proc. of ICDM '18, IEEE, 2018, pp. 1128–1133.

[25] Y. Elazar, Y. Goldberg, Adversarial removal of demographic attributes from text data, in: Proc. of EMNLP '18, 2018, pp. 11–21.

[26] P. Sen, D. Ganguly, Towards socially responsible ai: Cognitive bias-aware multi-objective learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 2685–2692.

[27] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proc. of NIPS '17, 2017, pp. 4765–4774.

[28] J. Chen, L. Song, M. Wainwright, M. Jordan, Learning to explain: An information-theoretic perspective on model interpretation, in: Proc. of ICML '18, 2018, pp. 883–892.

[29] S. Gao, G. Ver Steeg, A. Galstyan, Variational information maximization for feature selection, in: Proc. of NIPS '16, 2016, pp. 487–495.

[30] H. Lim, D.-W. Kim, Pairwise dependence-based unsupervised feature selection, Pattern Recognition 111 (2021) 107663.

[31] P. Zhou, L. Du, X. Li, Y.-D. Shen, Y. Qian, Unsupervised feature selection with adaptive multiple graph learning, Pattern Recognition 105 (2020) 107375.

[32] E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, in: ICLR (Poster), OpenReview.net, 2017.

[33] K. Chaudhari, A. Thakkar, Survey on handwriting-based personality trait identification, Expert Systems with Applications 124 (2019) 282 – 308.

[34] D. Castro, J. Tan, B. Kainz, E. Konukoglu, B. Glocker, Morpho-mnist: Quantitative assessment and diagnostics for representation learning, JMLR 20.

[35] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific Data 5.

[36] D. Hovy, A. Johannsen, A. Søgaard, User review sites as a resource for large-scale sociolinguistic studies, in: Proc. of WWW '15, 2015, pp. 452–461.