Computer Speech and Language 90 (2025) 101746

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

You Are What You Write: Author re-identification privacy attacks in the era of pre-trained language models

Richard Plant [a,*], Valerio Giuffrida [b], Dimitra Gkatzia [a]

[a] Edinburgh Napier University, 10 Colinton Road, Edinburgh, EH10 5DT, UK
[b] University of Nottingham, University Park, Nottingham, NG7 2RD, UK

ARTICLE INFO

Keywords:
Language models
Privacy-preserving
Differential privacy
Adversarial learning
Re-identification attacks

ABSTRACT

The widespread use of pre-trained language models has revolutionised knowledge transfer in natural language processing tasks. However, there is a concern regarding potential breaches of user trust due to the risk of re-identification attacks, where malicious users could extract Personally Identifiable Information (PII) from other datasets. To assess the extent of extractable personal information on popular pre-trained models, we conduct the first wide coverage evaluation and comparison of state-of-the-art privacy-preserving algorithms on a large multi-lingual dataset for sentiment analysis annotated with demographic information (including location, age, and gender). Our results suggest a link between model complexity, pre-training data volume, and the efficacy of privacy-preserving embeddings. We found that privacy-preserving methods demonstrate greater effectiveness when applied to larger and more complex models, with improvements exceeding >20% over non-private baselines. Additionally, we observe that local differential privacy imposes serious performance penalties of ≈20% in our test setting, which can be mitigated using hybrid or metric-DP techniques.

1. Introduction

Pre-trained language models (PTLMs), which can extend into the billions of parameters (Narayanan et al., 2021; Vaswani et al., 2017; Devlin et al., 2019), have driven many recent advances in the field of Natural Language Processing (NLP). These models are trained on very large textual corpora, contributing to state-of-the-art performance in numerous tasks and benchmarks (Han et al., 2021), and enabling the transfer of embedded knowledge to researchers and organisations lacking the resources to train an equivalently complex model (Qi et al., 2018; Peters et al., 2019).

However, PTLMs pose a threat to personal privacy because the more they learn about language from texts authored by real people (Petroni et al., 2019) and the better the models become at generalising beyond narrow task-specific use cases (Brown et al., 2020), the more they unavoidably reveal about the people who wrote them. A growing body of literature has demonstrated that PTLMs can leak substantial personally identifying information (PII) (Song et al., 2017; Carlini et al., 2019), thus offering the potential for misuse, such as re-identification attacks (Frankowski et al., 2006) or authorship attribution, revealing personal information about an individual, including potentially sensitive demographics (Rao et al., 2000; Emmery et al., 2021), which can harm public trust (Anwar, 2021; Shadbolt et al., 2019; Horvitz and Mulligan, 2015; Kozyreva et al., 2021; Prabhumoye et al., 2021). This work adopts the framework of the re-identification attack (Henriksen-Bulmer and Jeary, 2016), in which an adversary attempts to put together publicly-accessible information produced by or about an individual with secondary information sourced elsewhere

* Corresponding author.
E-mail address: r.plant@napier.ac.uk (R. Plant).

https://doi.org/10.1016/j.csl.2024.101746
Received 23 March 2023; Received in revised form 7 October 2024; Accepted 28 October 2024

Available online 16 November 2024

to illicitly recover private information about them. In our framing, the public data is text written by a person, something billions around the globe do as a matter of course–hence the title of this work, "You Are What You Write".

While there is little quantitative evidence currently available about the real prevalence and adoption of PTLMs in downstream applications, we can instead look at some alternate indicators: on a single popular website, BERT has been downloaded more than 24.3 million times,[1] while GPT-2 is the second most-downloaded at more than 17.5 million. This popularity represents a new risk factor for private information exposure available with minimal technical cost of entry, even when the task performance of the models themselves is outpaced by larger and more capable developments in the field (Brown et al., 2020; Narayanan et al., 2021; Chowdhery et al., 2022).

These privacy risks may facilitate the leak of PII, due to the memorisation of large parts of their training corpus (Carlini et al., 2019; Leino and Fredrikson, 2020) and the exposure of sensitive attributes through the embedding generation process (Song and Raghunathan, 2020), which can lead to exposure of political ideology (Iyyer et al., 2014; Colleoni et al., 2014), economic status (Aletras and Chamberlain, 2018; Doi et al., 2020), or unique identification (Narayanan and Shmatikov, 2008; Sun et al., 2012).

These privacy considerations become only more critical when viewed in the light of **under-resourced language** users, who already deal with exclusion and bias (Benjamin, 2018; Nekoto et al., 2020). Privacy differentials across multilingual representations is a relatively under-studied phenomenon, but we consider the potential existence of differences germane, and hence we extend our analysis to monolingual models for languages other than English, as well as multilingual models. There is some evidence that multilinguality can benefit task performance when applied to low or medium resource languages (Přibáň and Steinberger, 2021; Pires et al., 2019). However this is contested by other studies which claim a performance advantage for monolingual models (Rönnqvist et al., 2019; Virtanen et al., 2019). One possible explanation for this difference in findings is that the level of relatedness between high and low resource languages incorporated into a multilingual embedding may be determinative (Woller et al., 2021); if the lower-resourced language input is less related to the handful of high resource languages typically extant in language model training sets (Joshi et al., 2020), then the more complex multilingual model may underperform a smaller monolingual system. We wish to extend this analysis to consider the privacy implications of both strategies.

To reduce the impact of privacy leaks and attacks, we employ **adversarial training methods** (Plant et al., 2021; Goodfellow et al., 2015) which attempt to maximise the performance of a model when training and testing instances are drawn from similar but distinct distributions, by learning an intermediate representation that promotes features that benefit the target task, while suppressing features that are heavily conditioned by the domain shift. Applying this paradigm to privacy in NLP tasks is a short conceptual step: instead of domain-invariant features, we seek to surface private-attribute invariance while learning a suitable representation for our downstream task (Elazar and Goldberg, 2018). Alternatively, *Invariant Risk Minimisation* (Arjovsky et al., 2019) may be utilised in lieu of adversarial training to minimise personal information leakage. However, it has been found unstable under certain conditions (Rosenfeld et al., 2021).

To this end, we proposed CAPE, Context-Aware Private Embedding, a paradigm for training privacy-preserving models. In this paper, we further verify our paradigm with two new privacy-preserving techniques: Metric Differential Privacy (MDP) and Cross-Gradient Training (CGT), as well as a wider range of embeddings derived from different PTLMs.

To provide a **quantifiable guarantee** of disclosure risk, we employ the concept of differential privacy (DP) (Dwork, 2006). A computation is differentially private if the results on a dataset A are equally plausible as the results on a dataset B that differs in a single record, with the addition of a small amount of calibrated noise as described in Eq. (1). The maximum possible deviation, expressed by the parameter ϵ, stands in for our maximal privacy loss. The level of private information leaked by a computation M can be expressed by the variable ϵ, where for the symmetric difference $\oplus$ between two datasets A and B that differ in only one record, and any set of possible outputs $S \subset Range(M)$,

$$Pr[M(A) \in S] \leq Pr[M(B) \in S] \times exp(\epsilon \times |A \oplus B|) \tag{1}$$

We apply this to our tasks by adopting the local differential privacy approach (Cormode et al., 2018; Mahawaga Arachchige et al., 2020). Under this scheme, we apply noise proportional to the sensitivity of our computation – the maximum difference in output of the same operation carried out on both datasets – and the ϵ variable to the input before we carry out learning through gradient descent. This means we are essentially reducing the certainty of our model about the state of any individual record, and hence the ability of adversaries to carry out re-identification with intermediate representations that follow this step.

This work further consolidates CAPE (Plant et al., 2021), an adversarial/differentially private hybrid model as the preferred choice for achieving relatively high levels of individual privacy, while offering a significant level of calibration sensitivity so that downstream users can rapidly achieve an acceptable privacy/utility balance. On this basis, we investigate the impact of multiple task-related factors such as input dimensionality and model complexity on optimal privacy system design across languages. Specifically, we make the following contributions:

- Building upon our previous work (Plant et al., 2021), we conduct a systematic evaluation of privacy-preserving models, expanding our evaluation suite to include Metric-Differential Privacy (MDP) and Cross-Gradient Training (CGT) to further consolidate our proposed approach.

[1] https://huggingface.co/models.

- We provide a detailed comparison of the impact of such privacy-preserving approaches across a variety of embeddings derived from popular language models, thereby representing the largest proportion of potential risk. This demonstrates a reduction of up to 40% in relative attack efficacy.
- Additionally, we present results across a multilingual corpus to illustrate the effect of these technologies on a range of European languages, including relatively high-resource languages such as English, medium-resource languages like French and German, and lower-resource languages such as Norwegian and Danish. We also include a multilingual slice composed of records from all included languages. We identify both commonalities and distinctions in our results, which may offer useful areas for further linguistic study. To the best of our knowledge, this is the first study to extend this analysis beyond English.

This paper will first present the related work in Section 2, including some proposed methods for privatising texts in NLP. Then, we will lay out our methodology, models studied, and evaluation criteria in Section 3, before discussing the results in Section 4. Section 5 will present further analysis and discussion of secondary considerations arising from our experimental work. We will make some closing remarks in Section 6, along with recommendations for further work.

2. Related work

Differential privacy in NLP: Local differential privacy (LDP) has extensively been used in privacy-preserving data collection (Erlingsson et al., 2014; Kairouz et al., 2016; Wang et al., 2016). Furthermore, previous research into LDP has also borne out the potential for this method to reduce information leakage from trained text representations (Beigi et al., 2019; Sousa and Kern, 2022). Lyu et al. (2020) propose a system that adds perturbation to a pre-trained embedding vector by normalising the vector (Shokri and Shmatikov, 2015), then applying additive noise drawn from the Laplace distribution in the traditional DP fashion (Dwork et al., 2006).

Metric differential privacy (MDP) (Chatzikokolakis et al., 2013) generalises DP, involving the substitution of the original formulation's Hamming distance metric for understanding indistinguishably of two datasets for an arbitrary distance mechanism. Feyisetan et al. (2020) and Fernandes et al. (2019) applied this formulation to perturbing texts by proposing a system which swaps words for perturbed versions by considering the distance between their embedding vector with additive noise applied and the other vectors in the vocabulary, choosing the nearest candidate based on their metric of interest. Beyond simple distance metrics, some research show the potential value of alternate metric spaces (Feyisetan et al., 2019; Xu et al., 2020; Dhingra et al., 2018) in maintaining semantic continuity while ensuring DP-compliance.

Adversarial learning: On the other side, adversarial training has seen strong research interest in privacy-preserving applications (Zhang et al., 2020; Wang et al., 2019; Wang and Deng, 2018). Coavoux et al. (2018) applied this paradigm to personal privacy in the setting they describe as 'multi-detasking', in which an adversary network is trained using intermediate representations from a main task model to recover demographic labels about the input.

Li et al. (2018) applied adversarial learning to multiple protected attributes simultaneously across several English-language datasets, an approach that was also followed by Madras et al. (2018) and Elazar and Goldberg (2018). Xu et al. (2019) proposed a counter-intuitive model wherein input texts are rewritten via back-translation from adversarially-trained representations to eliminate sensitive attributes from the dataset. While this may prove a vital technology for dataset owners to apply before release, here we are more focused on the intermediate representations of the data.

This form of adversarial training has also been applied to debiasing representations: in Kaneko and Bollegala (2019), Zhang et al. (2018), and Zhao et al. (2018) the goal is to generate gender-invariant representations that do not express a differential between binary genders. Jaiswal and Provost (2020) and Li et al. (2020a) extend the problem space to include multi-modal data, an interesting research direction that may show potential for generalisation beyond NLP. In the medical field, research has focused on generative adversarial networks (GANs) for anonymising both imaging and genomics data (Bae et al., 2020; Cai et al., 2021; Arora and Arora, 2022), as well as the use of Invariant Risk Minimization for identifying and privatising invariant, and therefore presumably causal, features (Arjovsky et al., 2019; Zare and Nguyen, 2022).

Hybrid approaches: Some previous work leverages both differential privacy and adversarial training approaches: Phan et al. (2019) proposed an approach which implements classical differential privacy in an adversarial learning paradigm, however, this work relies on adversarial objectives to promote robustness to adversarial samples rather than privacy. Alnasser et al. (2021) proposes a similar hybrid approach, applying both adversarial and local differential privacy to English language embeddings, but only for the pre-trained BERT language model. In CAPE, we proposed a hybrid adversarial and local differential privacy approach designed for maximally private outcomes against a known set of re-identification tasks, extensible to arbitrary model architectures (Plant et al., 2021).

3. Methodology

We outline our task setting, model setup and baseline for evaluating the impact of privacy-preserving methods on an attacker's ability to recover indicator variables from intermediate representations in NLP tasks. The Trustpilot dataset by Hovy et al. (2015) is utlised, which contains review texts, scores, and demographic information (see Section 3.5 & Appendix A) for more details.

3.1. Task setting

We define as our base task sentiment prediction, that is inferring a 5-point rating from the linked review text, i.e. where $X = \{x_1, \ldots, X_n\}$ denotes a set of N text sequences and $Y = \{y_1, \ldots, y_n\}$ denotes their score $\in \{1, \ldots, 5\}$, we wish to learn a function f such that for a single sequence x_i, $f(x_i) = y_i$. To determine how successful an adversary may be in recovering private information from the intermediate representations within our model, we define a secondary classification task with access to our model's internal state and demographic labels for each row in our dataset. As such, we add a new term $Z = \{z_1, \ldots, z_n\}$ for the categorical value of the private variable, and our classification function becomes $f(x_i) = y_i, z_i$.

3.2. Model setup

We set the model parameters according to those used in Beigi et al. (2019), with our sentiment classifier consisting of a 200-unit hidden layer connected to a softmax output layer. Input to the classifier is a text representation generated by a feature extraction component consisting of one of a set of PTLMs (for a list see Section 3.7) along with two densely-connected layers is given as input to the base classification model. No additional regularisation (e.g. dropout) is utilised during training. Categorical cross-entropy is used as loss function during training. For each of our private indicator variables (gender, country, and age rank), we create a separate classification model with identical makeup to the base classifier, which will simultaneously attempt to learn the values of private indicators from the input representation (see Fig. 1). Early stopping was employed to mitigate overtraining and minimise overfitting on the result set, with training halted if the base loss of the main classifier did not decrease for more than 5 epochs.

3.3. Baseline

For comparison, we also reproduce the auto-encoder-based feature extraction and classification model in Beigi et al. (2019), under which the output of a trained encoder is substituted for the sequence embedding drawn from a language model.

3.4. Privacy-preservation

Once we establish the performance of our classification network, both base task and simulated attacker, for each of our test sets of language models, we introduce our privacy-preserving methods. These can be sorted into adversarial, differentially-private, and hybrid methods, and are described in full detail in Section 3.6. Our model architecture remains static during experimentation, aside from the specific additional steps listed for each privacy method.

3.5. Datasets

For the purposes of these experiments, datasets were selected that shared a number of characteristics: they consist of user-generated texts with a sufficient level of labelling to support traditional supervised NLP task inference, while also possessing a consistent set of demographic labels which enable the evaluation of attacks aiming to recover those attributes at the author level.

The primary resource selected for initial experiments was the Trustpilot dataset collected by Hovy et al. (2015) in 2015, due to its multilingual nature and highly consistent level of demographic labelling across multiple author features. However, a group of secondary monolingual datasets were also selected for comparison. These are explored in more detail below.

Trustpilot dataset. The Trustpilot dataset consists of user review texts along with accompanying 5-point scale scores for hundreds of thousands of users of the online consumer reviews site. While Trustpilot operates across many countries, the researchers restricted their data collection to those countries with more than 250,000 users at the time of collection: the United Kingdom, the United States, Germany, France, and Denmark. These countries represent significant linguistic variety since the user is not constrained to write their review in the prevailing language of the country in which they are posting.

Demographic variables are also included in this dataset. Users can choose to add their gender, birth year, and place of residence when posting their review. The researchers also partially augment these annotations by inferring gender from the existing national distribution of first names, and by using geographical databases to convert the free text location field into a latitude/longitude pair. For a longer description and some indicative statistics, consult Appendix A.

We split the dataset by language of review text rather than by country of posting, leaving us with six languages of interest with significant representation within the dataset: English, French, German, Danish, and Norwegian. We also generate a 50,000-instance multilingual set equally composed of texts from each listed language. We retain the country and gender annotations from the original dataset as our location and gender indicator variables, while birth years are used to sort the writers into three age rank bins (younger than 36, 36–45, older than 46) according to their age in 2021. We treat these features as categorical labels, converting them to an integer representation and applying one-hot encoding for use in learning. Further information on processing the data can be found in Plant et al. (2021), which describes our preliminary set of experiments using only a single set of pre-trained embeddings and English language instances from the dataset.

Splits are calculated using the adversarial Wasserstein process proposed by Søgaard et al. (2021): the input texts are arranged in Wasserstein space, a random centroid is selected and the nearest neighbours form the new test set. We use a 70/10/20 train/validation/test split structure, with the number of instances in each split shown in Table A.17 in Appendix A. These dataset splits were calculated before experimentation with the generated embeddings and hence the texts in each split are the same for every model tested.

Table 1
Details of additional datasets.

Name	Language	Task	Train size	Test size
Dioptra-L	English	Sentiment analysis	10,000	10,000
IberEval2017	Spanish	Stance detection	4319	1081
TAG-it	Italian	Topic modelling	10,000	9243
CLiPS	Dutch	Deception detection	1038	260
NoReC_gender	Norwegian	Sentiment analysis	3394	430

Table 2
Overview of the privacy strategies with their description.

Abbreviation	Description
Baseline (AE)	Auto-encoder embeddings w/ non-private classification
Baseline (FX)	Pre-trained embeddings w/ non-private classification
GR	Gradient reversal
CGT	Cross-gradient training
LDP	Local differential privacy
MDP	Metric differential privacy
CAPE	Context-Aware Private Embeddings

Secondary datasets. Since we were unable to source any other datasets which encompass the same breadth of languages and demographic labels as the Trustpilot dataset, we instead opted to select monolingual datasets sharing certain features of that set as a basis for comparison.

Our criteria for selecting appropriate datasets were as follows: we require a set of distinct NLP tasks which are demographic-independent, the data for which are also annotated with a private demographic variable which we are interested in measuring the leakage for. We also required that there be a source for pre-trained embeddings stemming from PTLMs available for those languages. This set of search parameters led us to the following datasets, the details of which are summarised in Table 1:

- Dioptra-L: an English-language sentiment analysis corpus (Kotze et al., 2021)
- IberEval2017: a Spanish stance detection corpus (Taulé et al., 2017)
- TAG-it: an Italian topic modelling corpus (Cimino et al., 2020)
- CLiPS Stylometry Investigation: a Dutch stylometry corpus (Verhoeven and Daelemans, 2014)
- NoReC_gender: a Norwegian sentiment analysis corpus (Touileb et al., 2020)

We discovered that the most productive demographic annotation for our purpose was gender; we were unable to discover any extant datasets with the same breadth of annotation as the previously-used Trustpilot dataset, and hence our new experiments use only the gender annotation.

In evaluating the previously-tested set of privacy strategies, we used the standard test/train splits established by the dataset owner, except where those did not exist, where we used a 20% random split method. Splits were limited to a maximum of 10,000 rows for resource efficiency. Embeddings for each language/task set were extracted from the 'xlm-roberta-base' multilingual model, selected for the languages in our experimental group using the same process used to select the first group of monolingual models in previous experiments. We were unable to compute MDP embeddings for this group given time constraints, and so results for this strategy do not appear in these results.

3.6. Privacy strategies

In this section, we describe the privacy-preserving techniques we applied in our experimental testing, as seen in Table 2.

Baselines: We adopted two baselines to benchmark the other private approaches. The first baseline is an auto-encoder method (AE) based on the work of Beigi et al. (2019), in which the model learns a document encoding in an initial training step before classification. The second uses a feature extractor component (FX) based on the work of Guo et al. (2019), in which the input to our network is the pre-trained sequence embedding drawn from the language model, which is passed through two 64-unit densely connected layers. In both models, the output of which is connected to two classifier heads, one for the base task and one for the simulated attacker. Each classifier head consists of a single 200-unit dense layer connected to a softmax output layer, as shown in Fig. 1. The loss function of such a model, taking as input a sequence encoding x_e and parameterised by θ, is shown in Eq. (2).

$$\mathcal{L}(x_e, y, z; \theta) = -log P(y|x_e; \theta) \tag{2}$$

Gradient Reversal: We define the adversarial objective as the performance of our secondary classifier in recovering the label of our private variable from the intermediate representation, scored by categorical cross-entropy, leveraging the *multi-detasking* idea as in Coavoux et al. (2018).

The loss function of a classification model parameterised by the variable θ_c can be combined with an adversarial classifier parameterised by θ_a to create an attribute-invariant model as shown in Eq. (3).

$$\mathcal{L}(x_e, y, z; \theta_c, \theta_a) = -log P(y|x_e; \theta_c) - \lambda log P(\neg z|x_e; \theta_a) \tag{3}$$

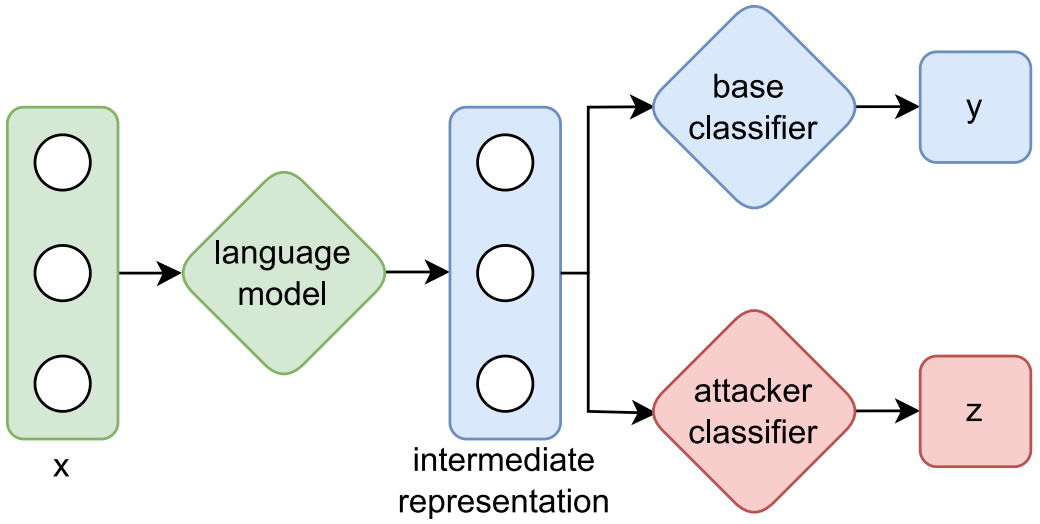

Fig. 1. Indicative model architecture for base and simulated attacker tasks, where x is the input text, y is the review rating score, and z is the private attribute label.

where x_e is an example from our training set, y is the target prediction label, z is the label of the desired invariant feature, λ is a regularisation parameter.

Cross-Gradient Training: CGT is an alternate formulation of the adversarial training as described in the previous section, as proposed in Shankar et al. (2018). The goal of CGT is not to suppress the signal that allows the secondary classifier to recover domain information entirely, since this may remove useful information that may be germane to our network's basic task. Instead, CGT involves an attempt to augment the input with a set of perturbations that allow it to generalise better to unseen domains, learned from the gradient of loss within the secondary classifier. This notion extends to privacy in the same sense as the previous example: if we can learn a generalised version of the private information that does not directly identify the demographic label, we can achieve an acceptable level of privacy.

Both strategies behave similarly during model training, as outlined in Algorithm 1. For a joint classifier parameterised by θ, where θ_c and $\mathcal{L}_c$ represent the parameterisation and loss function of the main task and θ_a and $\mathcal{L}_a$ represent the parameterisation and loss function of the adversarial task, ϵ represents a stability constant, α represents a ratio controlling the effect of the procedure on training, and where $x \in X$ represents the input data, $y \in Y$ represents the main task label, and $z \in Z$ represents the adversarial task label, parameter updates can be obtained,

Algorithm 1 Adversarial training procedure

1: Initialise θ_c and θ_a
2: Sample labelled batch (X, Y, Z)
3: $X_a := X + \epsilon \cdot \nabla_X \mathcal{L}_a(X, Z; \theta_a)$
4: $X_c := X + \epsilon \cdot \nabla_X \mathcal{L}_c(X, Y; \theta_c)$
5: $\theta_a \leftarrow \theta_a - \eta \nabla_{\theta_a}((1-\alpha)\mathcal{L}_a(X, Y; \theta_a) + \alpha \mathcal{L}_a(X_a, Y; \theta_a))$
6: $\theta_c \leftarrow \theta_c - \eta \nabla_{\theta_c}((1-\alpha)\mathcal{L}_c(X, Z; \theta_c) + \alpha \mathcal{L}_c(X_c, Z; \theta_c))$

Local Differential Privacy: LDP applies perturbations to the representation generated by the pre-trained language models. This technique was described by Lyu et al. (2020), aiming at reducing the certainty of a potential attacker about the true value of any recovered information, since they cannot reliably determine its difference from the other records in the dataset – what is often referred to in the literature as *indistinguishability* (Chatzikokolakis et al., 2015; Jorgensen et al., 2015). Converting the embeddings retrieved from our language model into a DP-compliant representation requires us to inject calibrated Laplace noise into the hidden state vector obtained from the pre-trained language model as $\tilde{x}_e = x_e + n$ where n is a vector of equal length to x_e containing i.i.d. random variables sampled from the Laplace distribution centred around 0 with a scale defined by $\frac{\Delta f}{\epsilon}$, where ϵ is the privacy budget parameter and Δf is the sensitivity of our function.

Since determining the sensitivity of an unbounded embedding function is practically infeasible, we initially followed the work of Lyu et al. (2020) in constraining the range of our representation to [0,1], as recommended by Shokri and Shmatikov (2015). However, as pointed out by Maheshwari et al. (2022), this version of the algorithm fundamentally underestimates the real sensitivity. We follow the corrected methodology, normalising the representation to ensure that the ℓ_1 norm and the sensitivity of our function

summed across n dimensions of x_e are the same, i.e. $\Delta f = 2$. LDP training procedure is demonstrated in Algorithm 2. For a classifier parameterised by θ trained using negative log-likelihood loss on a batch of examples with input data $x \in X$ and label $y \in Y$,

Algorithm 2 Local differential privacy training procedure

for $x \in X$ **do**
  Extract features from input batch: $x_e = f(x)$
  Normalise representation: $x_e \leftarrow x_e / \|x_e\|_1$
  Apply perturbation: $\tilde{x}_e = x_e + Lap(\frac{\Delta f}{\epsilon})$
  Add to perturbed batch: $\tilde{X} \leftarrow x_e$
end for
Train classifier: $\mathcal{L}(\tilde{X}, Y; \theta) = -log P(Y|\tilde{X}; \theta)$

Metric Differential Privacy: We also investigate the potential of MDP (Chatzikokolakis et al., 2013) through an implementation of the work of Feyisetan et al. (2020). With this approach, we continue to add calibrated noise to our representation. Unlike LDP however, the derived noise is not additive to the direct input to our learning mechanism, we instead use a metric-based system to find existing word vectors that are close by in the embedding space to our perturbed vector.[2] In the implementation we have chosen, a noise vector v_n is drawn from a multivariate normal distribution parameterised by the dimensionality of the embedding vector and constrained to within the unit ball. This vector is then scaled by sampling magnitude values l from the Gamma (or maximum entropy) distribution, constrained by the privacy parameter ϵ. We then add our scaled noisy vector lv_n to the original vector v_i, and use a convenient distance metric to find the nearest existing vector within our PTLM vocabulary: in the case of our implementation as in the original work, it is the Euclidean distance.

The procedure for extracting a perturbed embedding vector is demonstrated in Algorithm 3. For an input sequence $x = \{w_1, \dots, w_n\} \in X$, language model φ with existing dictionary D, and privacy parameter ϵ,

Algorithm 3 Metric differential privacy training procedure

for $i \in 1, ..., n$ **do**
  Get embedding vector $v_i = \varphi(w_i)$
  Sample noise vector from multivariate normal distribution $v_n = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$, where m is the dimensionality of the embedding, Σ is the covariance matrix set as identity, and the mean μ is centred at the origin
  Draw a sample l from the Gamma distribution $l = \frac{x^{m-1}e^{-x/\phi}}{\Gamma(m)\phi^m}$, where $\phi = 1/\epsilon$
  Scale the noise vector as $v_n = v_n * l$
  Perturb embedding $\hat{v}_i = v_i + v_n$
  Get the nearest word to perturbed vector $\hat{w}_i = argmin_{d \in D} \| \varphi(d) - \hat{v}_i \|$
  Swap x_i for $\hat{w}_i$
end for

Context-Aware Private Embeddings: To preserve the general privacy benefits of DP-compliant embeddings with invariance to the specific private variable identified for adversarial training, we combine both processes in a model called Context-Aware Private Embeddings (CAPE) (Plant et al., 2021). We add calibrated noise drawn from the Laplace distribution to the pre-trained sequence embedding obtained from the language model, as in Lyu et al. (2020), as well as using the joint loss function with gradient reversal stemming from Coavoux et al. (2018). A diagram of the combined system is pictured in Fig. 2.

In Section 4.2, we report the performance of our attacker classifier when we apply our privacy-preserving models to each language dataset individually, as well as the multilingual set. Gradient reversal (GR) and cross-gradient training (CGT) methods are applied as previously described: for GR, we simply invert the sign of the gradient during backpropagation. CGT requires us to watch the gradient of the attacker classifier during training to perturb the input to both classifiers. We set the step size of perturbation during training to 5 and the α parameter, which controls the ratio of the contribution of the perturbed input versus the original input to the total loss, to 0.5.

In terms of our locally differentially private models, both LDP and CAPE, we add calibrated noise from the Laplace distribution to our sequence embedding between the output of our language model and the first dense layer of our network. The ϵ parameter controlling the size of our privacy budget is kept static at 0.1 throughout this stage of testing. To test metric differential privacy (MDP), we instead pre-compute all sequence swaps across our dataset before training or inference, since exhaustively searching the embedding space for every word vector during online training proved highly computationally expensive. In this case, the ϵ privacy parameter is set to a value of 20 during testing. Base and adversary classifiers are trained and tested with noised sequences obtained in these ways in the same fashion as the other experimental models.

[2] Note that dependent on the magnitude of the noise added to the sequence and the size of the dictionary of pre-trained embeddings, the closest existing representation in embedding space to the noised word may be the original word.

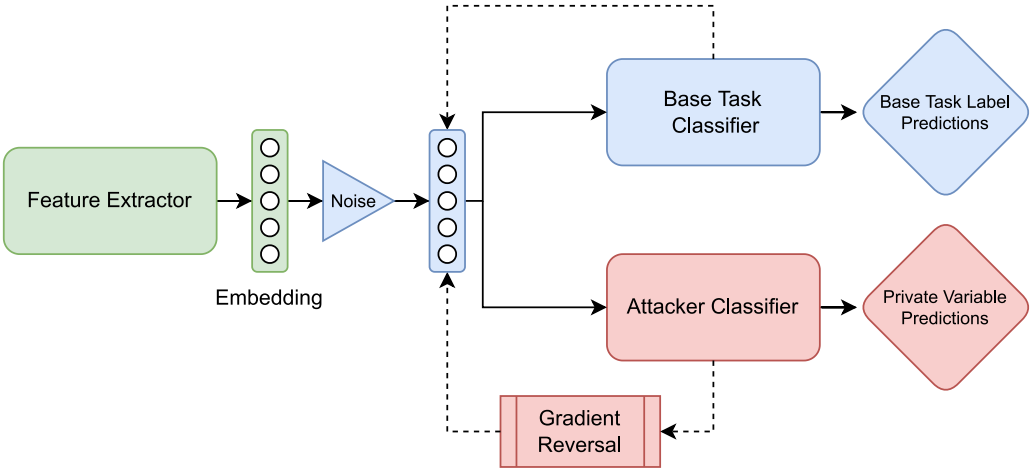

Fig. 2. CAPE model diagram. Dashed lines represent gradient updates.

3.7. Language models

We present here the language models chosen to evaluate as a source for pre-trained text representations which may present a privacy risk. A full description of each chosen embedding source can be seen in Appendix B.

Language models were chosen based on the number of downloads listed on the Huggingface model repository website,[3] which we consider a reasonable proxy for popularity/impact. We ranked the top five models for the English language and attempted to find matching models that have been developed for each of the other languages of interest. This process led us to select models based on BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and GPT-2 (Radford et al., 2018). We also add GloVe (Pennington et al., 2014) and Komninos (Komninos and Manandhar, 2016) embeddings for English, as representatives of previous generations of pre-trained vectors versus the more recently popular language models. This kind of parity was more difficult to achieve for multilingual models, given that successful monolingual models are often not transferred for classification tasks across languages, and hence we include the 'paraphrase-xlm-r-multilingual-v1' (Reimers and Gurevych, 2020) and 'distiluse-base-multilingual-cased-v2' (Yang et al., 2020) models.

Consistency in model architecture was not a goal of the selection process, since the objective was primarily to evaluate the risk level represented by re-identification attacks on popular model outputs. Thus, care should be given in interpreting the results to take account of this fact, along with other peculiarities of each model as listed in Table B.19.

3.8. Evaluation

We consider the primary metric of success in our attacker task as the ability to reliably distinguish the gender, age, or location of the individual from the text representation. We establish the rate of successful predictions of our demographic attribute as a measure of the privacy at risk, as in Jaiswal and Provost (2020) and Zhao et al. (2019) – any change in this from our baseline to private methods will indicate this. Since we are dealing with an imbalanced dataset across multiple axes of comparison, we report both accuracy and F1-scores, emphasising the latter as a more reliable measure.

To measure statistical significance when comparing more than two models, we provide summary statistics for each experiment using the two-way ANOVA method, reporting the variance in F1 results (σ^2) across the private information type categorical variable row-wise, and the summary sum of squares (SS), degrees of freedom(df), F-value (F), F critical value (F-crit) and p-value (P) scores both row- and across the embedding type column-wise, and across the whole set of results. F-crit here is calculated based on a 0.05 p-value significance threshold.

Since we rely on empirical demonstrations of our privacy outcomes, we also consider utility through the same lens. Performance in the downstream task, prediction of the review score class from the text as laid out in Section 3, forms our primary metric in comparison to the baseline model. We report F1-score as a robust mechanism for use across our test scenarios.

4. Results

In this section, we will discuss in detail the results of each experiment, as well as discuss some initial findings from our empirical data.

[3] https://huggingface.co/models?language=en&sort=downloads.

Table 3
Non-private task baseline comparison. Gl: GloVe, Ko: Komninos, Be: BERT, MB: Multilingual BERT, Ro: RoBERTa, Al: ALBERT, GPT: GPT-2.

Model	Private variable								
	Base task		Gender		Location		Age		σ^2
	acc.	f1	acc.	f1	acc.	f1	acc.	f1	
Baseline (FX)									
Gl	0.646	0.624	0.749	0.674	0.732	0.661	0.752	0.674	0.021
Ko	0.631	0.598	0.749	0.668	0.713	0.630	0.757	0.680	0.024
Be	0.680	0.670	0.788	0.737	0.794	0.749	0.794	0.746	0.027
MB	0.662	0.654	0.763	0.712	0.749	0.696	0.752	0.691	0.021
Ro	0.690	0.681	**0.802**	0.765	**0.811**	**0.771**	**0.811**	**0.773**	0.026
Al	**0.691**	**0.685**	0.795	**0.779**	0.806	0.765	0.801	0.754	0.029
GPT	0.689	0.681	0.792	0.758	0.789	0.730	0.756	0.753	0.024
Baseline (AE)	0.599	0.449	0.537	0.375	0.755	0.649	0.599	0.449	0.034
ANOVA metrics	SS	df	F	P	F-crit				
Row	0.069	7	0.783	0.613	2.764				
Column	0.236	2	9.318	0.003	3.739				
Total	0.482	23							

4.1. Non-private baselines

We present here the results from a comparison of our basic method to the auto-encoder-based method of Beigi et al. (2019). We use this baseline method in successive experiments to establish a non-private benchmark for private information leakage that we can compare to each of our privatising methods to quantify their impact. It should be noted that we have not attempted to implement the privacy-enhancing elements of Beigi's work, and therefore these results establish only a comparable level of performance for both the basic sentiment analysis task and the private information re-identification task for our system with a recent published baseline. Results for the base task, as well as each private demographic variable attacker network, can be seen in Table 3. Accuracy and F1-score are presented for each embedding class, with the best-performing instance, e.g. the classifier most able to predict the target variable, highlighted.

We apply these methods to only the English-language instances from our dataset, to more accurately represent a comparison with the original work, which included only English results from the Trustpilot set.

Several interesting trends can be observed immediately from inspection of the results. First, using pre-trained embeddings from popular PTLMs without significant fine-tuning immediately outperforms an auto-encoder-based approach trained only on instances from our dataset. This finding is perhaps unsurprising since the training corpus and depth of the model are far less extensive. These approaches may be eminently compatible (Lewis et al., 2020; Gordon et al., 2020; Li et al., 2020b).

Secondly, it appears that pre-trained embeddings that lead to higher performance on the basic sentiment analysis task also enable better performance for the attacker, as shown by the results for RoBERTa and Albert in Table 3. We posit that by becoming more efficient at encoding contextual semantic information through the co-occurrence of terms in their training sets, these models have also become more vulnerable to private data leakage through encoding elements of authorial style (Emmery et al., 2021). This effect is not negligible, with the largest relative change in F1-score amounting to $\sim 16\%$.

4.2. Privacy-preserving models

In this section, we present the results for each of our previously cited privacy-preserving processes of interest, as detailed in Section 3.6, applied across the various languages in our Trustpilot dataset. We investigate here not only each language corpus separately but a combined multilingual set drawn from each. For each privacy-preserving model, the set of attacker tasks aiming to predict the gender, location, and age variables are applied separately to each set of pre-trained representations.

Results for private variable classifiers are presented as F1-scores. Results in the table are presented at a precision of 4 digits. Best performance — that is, showing the poorest record of predicting the target variable — is picked out in bold. We present also the outcome of our utility testing, where the effect of increasing privacy on the performance of our model is measured as a function of the performance of the base task classifier—as the level of privacy rises, we would expect to see a concomitant drop in utility as a trade-off (Li and Li, 2009; Rastogi et al., 2007).

Results were produced for each embedding source, privacy strategy, and language group separately, the data for which were aggregated and grouped to produce the results tables, as shown in Tables 4, 5, and 6. Note: we group model types by their base structure, i.e. CamemBERT as a version of RoBERTa. These tables also include aggregated task performance as F1-scores - in this case, higher scores represent better model utility. However, caution is warranted since the data used to train each individual model will vary in extent and quality. Full results for each language are available in Appendix C.

We can derive several interesting observations from our English-language results. Firstly, it is clear that the application of both broad categories of privacy-preserving mechanism — differential privacy noise and adversarial training — produce a marked drop in the information leaked to our simulated attacker network. In fact, in the best case, that of RoBERTa embeddings passed through the CAPE model, the F1-score of our gender attacker experiences a relative drop of $\approx 45\%$ over baseline.

Table 4
Summary F1-score results produced across all runs of all source/language/privacy strategy scenarios, presented as aggregated results grouped by input language.

Language	Gender	Location	Age	Task
English	0.503	0.689	0.498	**0.609**
French	0.504	0.984	0.415	0.476
German	0.57	0.985	0.342	0.526
Danish	0.506	0.996	0.257	0.447
Norwegian	**0.463**	0.995	**0.237**	0.493
Multilingual	0.475	**0.628**	0.383	0.508

Table 5
Summary F1-score results produced across all runs of all source/language/privacy strategy scenarios, grouped by embedding source.

Embed source	Gender	Location	Age	Task
Glove[a]	0.499	0.670	0.507	0.593
Komninos[a]	0.498	0.658	0.509	0.598
ALBERT[a]	0.496	0.720	0.502	**0.641**
XLM[b]	0.501	**0.597**	0.378	0.548
USE[b]	**0.477**	0.539	0.386	0.501
BERT	0.516	0.930	**0.348**	0.530
Multi-BERT	0.490	0.903	0.352	0.487
RoBERTa	0.513	0.931	0.352	0.503
GPT-2	0.536	0.887	0.419	0.543

[a] English-only.
[b] Multilingual-only.

Table 6
Summary F1-score results produced across all runs of all source/language/privacy strategy scenarios, grouped by privacy strategy.

Privacy strategy	Gender	Location	Age	Task
Baseline	0.669	0.863	0.508	**0.602**
CGT	0.438	0.842	0.352	0.421
GR	**0.422**	0.827	0.319	0.588
LDP	0.449	0.863	0.360	0.462
MDP	0.631	0.913	0.409	0.583
CAPE	0.437	**0.821**	**0.307**	0.413

This drop is variable across the range of embeddings tested, but several tendencies can be observed in action. Firstly, models with more parameters show larger mean reductions in attacker performance. We note when comparing RoBERTa, ALBERT, BERT, and GPT-2, that the results for the change in performance of gender re-identification against the baseline appear to be proportional to the size of the model in terms of the number of parameters. ALBERT shows an ≈25% average drop in performance with the smallest set of 11 million parameters, BERT shows an ≈30% drop with 109 million, GPT-2 shows a reduction of ≈32% with 117 million, while RoBERTa shows the largest relative drop of ≈39% with 125 million parameters, as seen in Table 5.

It appears that adversarial methods produce remarkably homogeneous results. The variance between results for different sets of embeddings for our purely adversarial methods (gradient reversal and cross-gradient training), is much lower than that of our other privacy-preserving methods. Indeed, at the level of precision displayed in , the vast majority of scores are identical. We theorise that this effect is due to a decrease in variance across the embeddings caused by the suppression of contextual stylometric cues that also impact the existing contextual semantic relationships between token vectors.

Adding adversarial training objectives decreases information leakage in a DP-compliant scenario. Comparing the mean performance reduction over baseline for all private demographic variables across our differentially-private systems—local differential privacy, metric differential privacy, and CAPE–we find that CAPE provides an ≈24% reduction, while LDP and MDP show smaller drops of ≈18% and ≈4% respectively, as seen in Table 6. It is worth pointing out that the difference in MDP score is likely caused by the differing privacy budget parameters, and does not hold general validity beyond this specific experiment.

We note that in the case of our baseline and CAPE models, embeddings extracted from mBERT show anomalously low age identification results. One potential explanation for this outcome is the vocabulary peculiarities of mBERT; while the model vocabulary is much larger than BERT (119,000 vs. 30,000 entries), minority languages are oversampled in relation to English (Abdaoui et al., 2020), which could indicate that archaisms or neologisms that may indicate age are being left out.

French and German results: We claim that the language models used to generate embeddings for these languages represent a fair comparison to the English language models used to generate the previous set of results, since the training data corpus for each is comparable in size and scope. For instance, the RoBERTa-derived French model CamemBERT (Martin et al., 2020) was trained using 138 GB of French documents from the multi-lingual OSCAR dataset (Ortiz Suárez et al., 2019), which compares favourably to the mixed 160 GB of documents in the original model training corpus (Liu et al., 2019). This holds true also for FlauBERT (Le

et al., 2020) and German BERT (Chan et al., 2020) as compared to the original BERT (71 GB and 12 GB vs. 13 GB). On this basis, we believe that the set of models used in these tests can be considered as relatively even in terms of language resources, and hence we believe that any observed deviation of results can be attributed to changes in the language structure and textual content, rather than the effect of sparse training/fine-tuning text availability.

From our results, we see that re-identification of gender is higher on average than in English results. The mean average results across all privacy-preserving technologies is higher for both French (≈4% relative) and German (≈20% relative) when compared to English, as seen in Table 4. French and German exhibit more gendered language features than English (Hord, 2016; Kokovidis, 2015; Stahlberg et al., 2007).

However, the performance of age identification across all models is much lower. Our results show an absolute drop in age re-identification performance of ≈4% for French and ≈11% for German. We note that this may be due to the increased prevalence of fake data for this attribute linked to reduced willingness to share accurate personal information in these nations, as discussed in Section 1.

Location in terms of country now ceases to be a meaningful target variable. While testing with English-language instances we could expect significant proportions of source texts to be tagged with countries other than the UK, since Trustpilot also operates in numerous other nations where English is the primary language. For French or German, where we would expect the vast majority to be sourced from the country in question. In these cases, identifying the country of origin does not represent an increased privacy threat — the text itself does that reliably. Regional and locality identification would still present a threat, although one not addressed in this experiment.

Overall, we continue to note the reduced performance of attacker networks in the CAPE setting over other differentially-private models, which continues to demonstrate the effectiveness of adversarial training at reducing information leakage.

Danish and Norwegian results: We claim that these languages can be understood as representing a typical scenario for medium-resource languages (Ortiz Suárez et al., 2020; Iliev and Genov, 2012). In these cases, the language model has been pre-trained in the same way as the original research in a high-resource setting, but with a substantially reduced text corpus. The Danish BERT model used in these experiments for instance is trained on around 1.6 billion words of text, while the original BERT model was trained with around 3.3 billion words. This was also the case for the NorBERT model used to produce Norwegian embeddings, which was trained with around 2 billion words. In this way, we can expect the model to be less sensitive to context variation and less able to represent complex text sequences, since we would intuitively expect general task performance to scale with corpus size and variability (Sasano et al., 2009; Rose et al., 1997).

Gender re-identification performance is lower than higher-resource settings. Comparing the performance across all embeddings of our gender prediction task, we can see that both Danish and Norwegian show comparatively lower scores than any previously analysed language set. In terms of mean F1-score, Danish shows an ≈3% absolute drop in performance, while Norwegian shows a larger ≈6% absolute reduction, as seen in Table 4. This lower performance ratio is mirrored in the results once privacy-preservation is applied. This result could be linked to the lower sensitivity of the model due to limited training texts, as previously mentioned, although we also contend that ongoing changes in the contextual and grammatical markers of gender in online language use for these languages could also be a factor (Lohndal and Westergaard, 2021; Cornips and Gregersen, 2017; Rodina and Westergaard, 2015).

The accuracy of age classification is also lower than in all other experiments. We note that the performance of our re-identification network in the majority of settings is lower by an average of >10% for both lower-resource languages than the other language settings studied. This may reflect more public resistance to sharing age as mentioned previously, or a lack of variation in the training corpora; it could be that the Danish and Norwegian sections of Common Crawl and Wikipedia do not contain as much stylistic variation between authors of differing age ranges as the larger English or French sections.

Finally, we turn to the results for privacy models trained using a multi-lingual corpus. We can expect this model to show lower baseline performance and sensitivity overall than models trained using a large corpus of documents in the individual languages (Pires et al., 2019; Rust et al., 2021). However, we do establish by examining the baseline results that cross-lingual demographic prediction is viable given a multilingual embedding, indicating that such models are vulnerable to the same information leakage as monolingual language models.

Prediction at the national level across the set of languages in our dataset proves meaningful once more, given the heterogeneous set of localities in the test set. We note that certain applications of privacy-preserving systems here enhance the ability of our attacker network in specific settings—for instance when applying differential privacy across mBERT embeddings, or using cross-gradient training. We suggest that this effect may be due to the additional noise pushing the distinct clusters of single language embeddings further apart in a way which may make them easier for our attacker network to classify.

Gender results however trend lower than monolingual models, more closely resembling the Danish and Norwegian scenarios than the relatively higher-resource languages. We suggest that it may be that models with a larger and more diverse training corpus in a target language are better able to represent the gendered nature of textual sequences, and hence are more liable to leak information.

In assessing the utility outcomes of our privacy interventions, as illustrated in Figs. 3 and 4, we find that outcomes are better for higher-resourced languages. Averaged across all models, task performance for English outpaces all other language groups, exhibiting a maximum ≈15% absolute lead over the Danish group and minimum ≈9% absolute over the multilingual group. Norwegian and Danish results are also lower in absolute terms than French and German, supporting our intuition that this finding is relative to the degree of pre-training and the extent of the training corpus.

Gradient reversal and metric differential privacy methods both exhibit only around ≈2% drop in base performance in absolute terms, which dependent on the task in question may represent a negligible downstream effect. This finding is interesting given

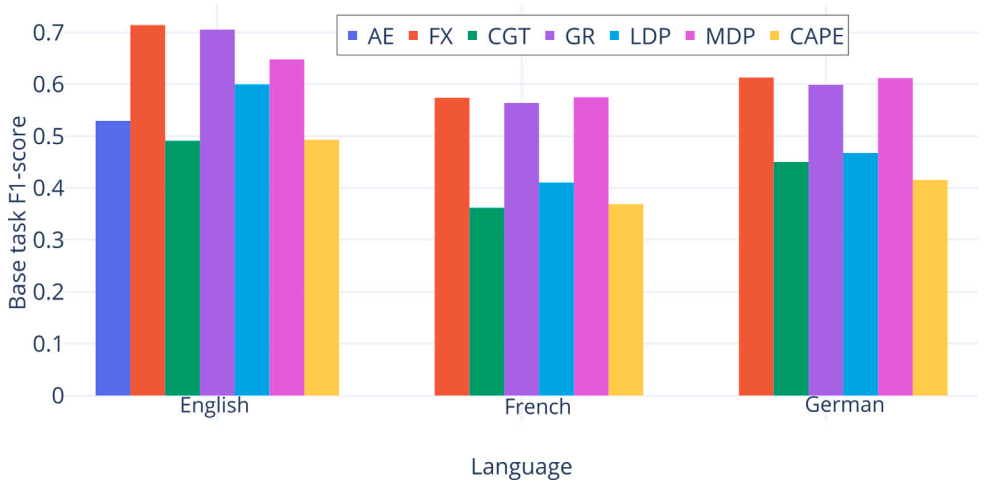

Fig. 3. Utility results for English, French, and German models.

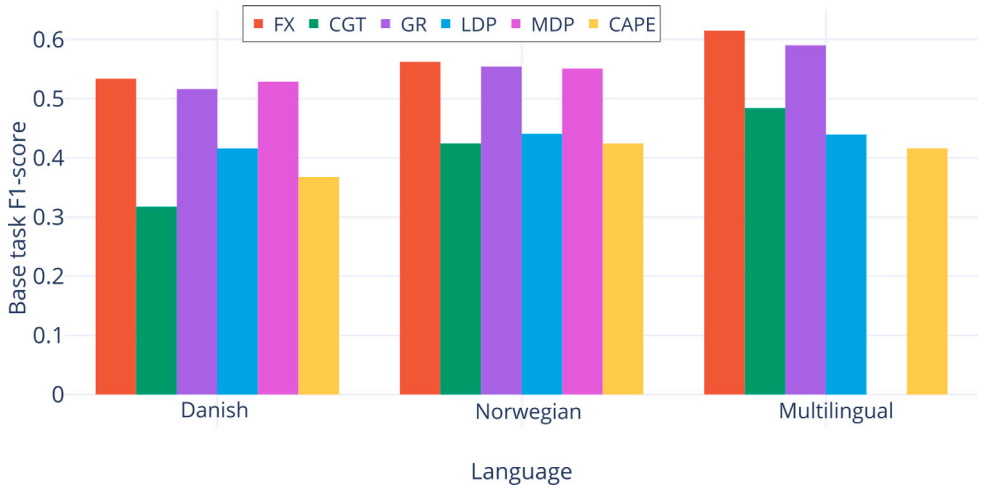

Fig. 4. Utility results for Danish, Norwegian, and Multilingual models.

the divergent levels of privacy offered by the respective methods: in English results for instance, GR exhibits a re-identification performance reduction of around ≈13% absolute. Given that MDP is a general privacy measure and does not require the introduction of a known set of demographic annotations, this is perhaps an explainable phenomenon. However, additional testing could reveal the extent of the general privacy guarantee; for more, see Section 6.1.

Local differential privacy however introduces a high level of utility perturbation at a fixed privacy budget. Both systems that rely on additive noise, LDP and CAPE, display a similar level of utility degradation at ≈20% absolute penalty, higher than the other systems addressed in our testing. This is to be expected since we have set a very restrictive privacy budget ($\epsilon = 0.1$) to achieve very high levels of privacy risk reduction in our experimentation.

4.3. Additional task settings

Since the previous experiments took place in a single task/dataset setting, we reproduced our methodology as far as practical across other extant NLP datasets in order to evaluate the general applicability of our methods, and especially whether the high privacy/low utility trade-off seen in the case of differentially-private strategies holds true in other scenarios.

The same methodology and evaluation criteria as in the previous experiments are applied across the experimental group, with reduction in attacker performance in determining the gender of the input sequence author as measured by F1 score the primary measure of success. We also include the results of a random classifier as an additional baseline measure. Results for each dataset/task are presented in Table 7 and Fig. 5.

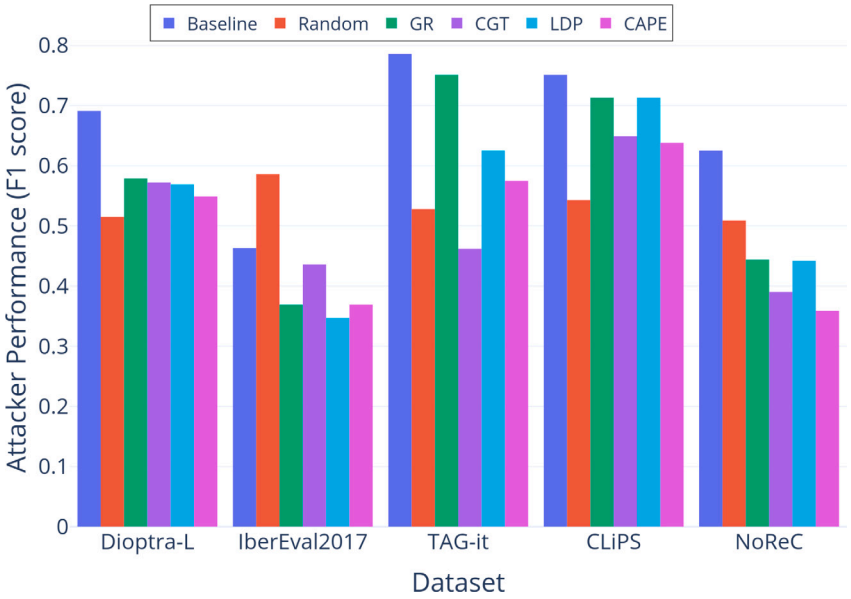

Fig. 5. Privacy strategy results for additional datasets.

Table 7
Results of privacy strategy experiments on additional datasets. Bold values represent best privacy strategy performance.

Dataset	Strategy	Task	Attacker	Dataset	Strategy	Task	Attacker
Dioptra-L	Baseline	0.342	0.691	CLiPS	Baseline	0.788	0.751
	Random	0.209	0.515		Random	0.488	0.543
	GR	0.354	0.579		GR	0.793	0.713
	CGT	0.379	0.572		CGT	0.713	0.649
	LDP	0.333	0.569		LDP	0.580	0.713
	CAPE	0.334	**0.549**		CAPE	0.510	**0.638**
IberEval2017	Baseline	0.659	0.463	NoReC	Baseline	0.167	0.625
	Random	0.365	0.586		Random	0.197	0.509
	GR	0.598	0.369		GR	0.189	0.444
	CGT	0.314	0.436		CGT	0.167	0.390
	LDP	0.436	**0.347**		LDP	0.117	0.442
	CAPE	0.436	0.369		CAPE	0.194	**0.359**
TAG-it	Baseline	0.479	0.786				
	Random	0.107	0.528				
	GR	0.468	0.751				
	CGT	0.651	**0.462**				
	LDP	0.458	0.625				
	CAPE	0.442	0.575				

Results in these experiments are broadly in line with the trends established in the Trustpilot dataset experiments: we note that CAPE provides excellent privacy performance in most tasks, while cross-gradient training is more task-specific, providing the best performance in only the TAG-it dataset.

Utility results are also in line with expectations, as we see both LDP and CAPE strategies provide the lowest task performance numbers, with the exception of the NoReC_gender dataset, where no strategy (including no privacy-preservation at all) provides task performance reaching the threshold of random guessing.

This enhances our findings in the previous experiments that consistently show differentially-private solutions providing the highest level of privacy preservation, while gradient-based privacy strategies are far less consistent and more task-dependent. We also note that CAPE provides superior preservation outcomes in every task setting aside from the IberEval2017 task, the cause for which deserves further investigation.

5. Discussion

This section will provide some further context and analysis for our results in the previous experiments, including the impact of some potential confounding factors on the level of privacy provided by our suggested systems.

Table 8
Attack effectiveness and change in score vs. non-private baseline across a range of BERT model sizes.

Model	Parameters (million)	Gender	Location	Age	vs. Baseline
Tiny	4.4	0.473	0.626	0.316	−23%
Mini	11.3	0.475	0.632	0.310	−25%
Small	29.1	0.477	0.651	0.305	−27%
Medium	41.7	0.480	0.598	0.307	−29%
Base	110	0.488	0.627	0.293	−30%

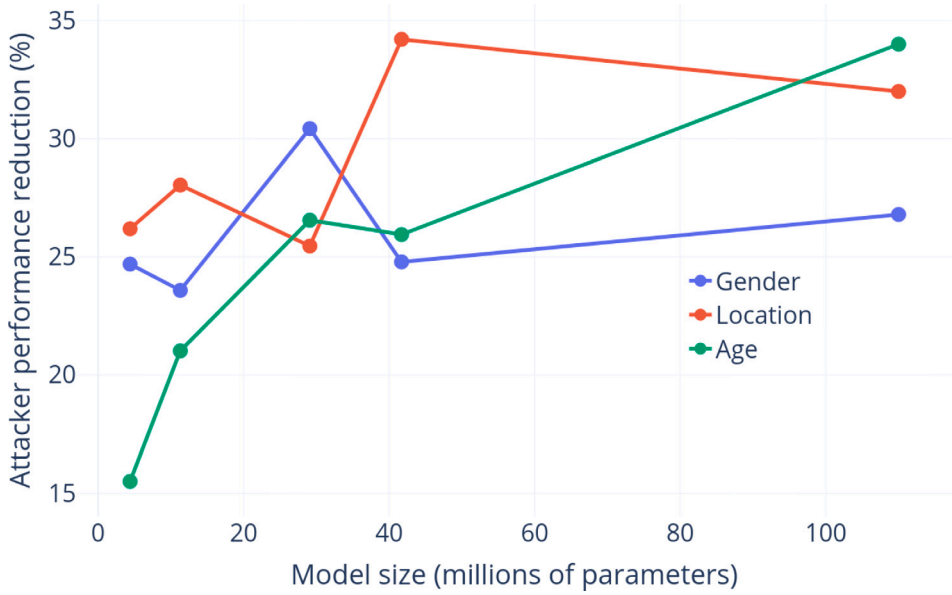

Fig. 6. Model size and reduction in attacker performance with privacy-preservation applied.

5.1. Examining the role of model size

In our discussion of the results of our experiments, we have indicated that we find some evidence of a trend towards models with more parameters showing more pronounced privacy-preserving outcomes when exposed to our set of interventions. However, our observation of this correlation could be somewhat spurious, an effect of confounding factors such as slight discrepancies in model setup or training data distribution. In order to provide a better basis for making such a claim, we must also perform a set of experiments with a group of models which share more commonalities, with their major distinction being the size of the model, as measured by the number of parameters they contain.

To this end, we adopt a range of distilled BERT models produced by Turc et al. (2019): Tiny, Mini, Small, and Medium (4.4M, 11.3M, 29.1M, and 41.7M parameters respectively) as well as the BERT-base model used in our other experimental work (110M parameters). We produce embeddings for a subset of 20,000 instances from the English and Danish parts of our dataset using this set of models and compare for each the relative performance of our privacy-preserving measures over information leakage measured via our baseline re-identification attack. We present the mean performance as F1-score of our attack in all scenarios per model size plus an indication of the average reduction ratio of attacker performance over the non-private baseline in Table 8.

While the trend in increasing information loss as model size increases is relatively evident when analysing the gender demographic variable, the picture is much more mixed when considering both location and age. While age attack results are poor across the board, performance is actually minimised when using the largest version of our model here. We can however note that the effect of privacy preserving strategies is consistently greater when using the largest versions of models versus the smallest versions, as illustrated in Fig. 6. More experimentation is required to determine the cause of this effect; it may be that larger models are more susceptible to attacks of this nature and privacy preservation is reducing this tendency, or it may be the inverse, smaller models are in fact more prone to leakage and privacy preservation is less effective in those settings.

5.2. Assessing embedding deviation

During our experiments, the question occurred of how to adjudge the potential effects on privacy and the utility of preservation techniques without the investment of extensive empirical research. Here, we propose to apply any potential technique which involves

Table 9
Average distance metrics for English text sequences paired with equivalents produced by LDP and MDP models.

	Euclidean	Cosine
LDP	387.987	0.987
MDP	6.017	0.168

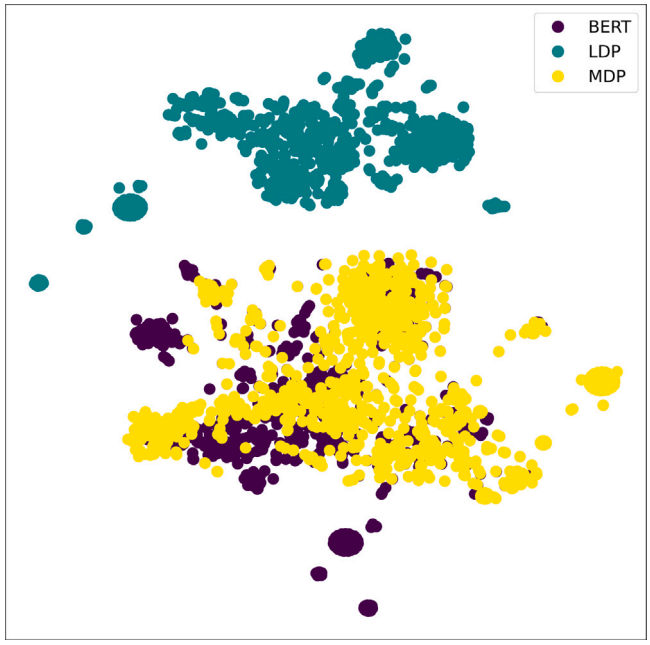

Fig. 7. Plot of embedding vectors for English text sequences produced by the 'bert-base-cased' model, and their local DP and metric DP equivalents. Dimensionality reduced using t-SNE.

Table 10
Average distance metrics for Danish text sequences paired with equivalents produced by LDP and MDP models.

	Euclidean	Cosine
LDP	387.830	0.935
MDP	2.940	0.036

altering the content of embeddings to a subset of the dataset, and then projecting the vectors into a lower-dimensional space where large-scale effects may be more obvious.

To test this, we gather a slice consisting of the first 10,000 rows in our English language corpus and extract the pre-trained embeddings from the *bert-base-cased* model. For comparison, we extract the same set of sequences from our metric differential privacy model, and from our local differentially private model. We present the average Euclidean and Cosine distance metrics (calculated for each paired row) in Table 9.

Our empirical results demonstrate the distinct effect of the differential privacy process, but the outcome may be easier to distinguish in a visual form. To project our dataset slice into a two-dimensional space, we adopt the t-SNE dimensionality reduction process (Maaten and Hinton, 2008), which attempts to minimise the KL-divergence between a similarity distribution over high-dimensional objects and a corresponding low-dimensional representation.

The results of the t-SNE dimensionality reduction process can be seen in Fig. 7, with the original embedding compared to the equivalent sequence with local and metric DP applied. We observe that the outcome of the dimensionality leads to a linearly separable grouping of local DP sequences, while the original and MDP results are far more contiguous.

In order to test whether this is an artefact of the particular dataset in use, or due to specific features of the language model, the experiment was re-run using Danish-language input sequences and embeddings derived from the *danish-bert-botxo* model. The remarkable homogeneity of MDP and BERT-derived results can also be seen in the results, shown in Table 10 and Fig. 8, where the average deviation is less than in English. Instances perturbed using the LDP process again show significant deviation in the t-SNE plot from the base embeddings and the MDP-perturbed results, which overlap to a very large extent. When compared to the English results in Fig. 7, similar trends in the separability of the instances can be observed. Since the t-SNE process is an attempt to minimise

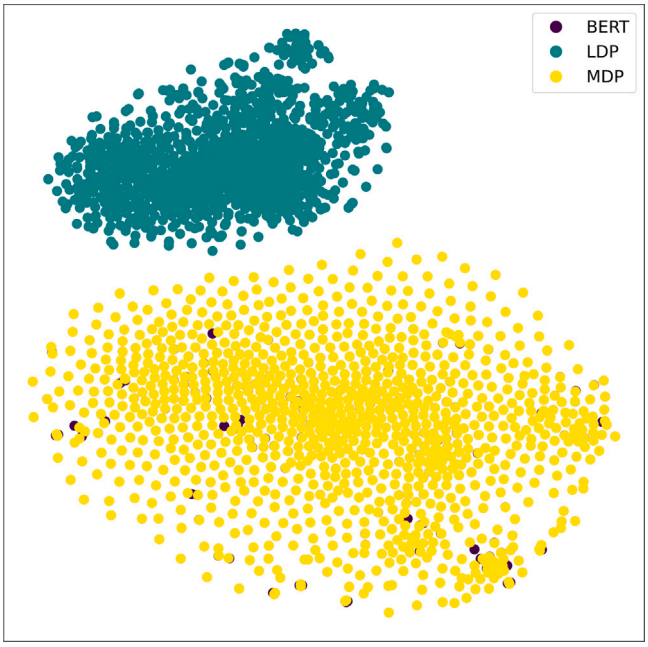

Fig. 8. Plot of embedding vectors for Danish text sequences produced by the 'danish-bert-botxo' model, and their local DP and metric DP equivalents. Dimensionality reduced using t-SNE.

the KL divergence between pairs of points in the projection, we would expect similar inputs to largely overlap, which we again fail to see for LDP inputs.

Intuitively, this supports our earlier results - the LDP-perturbed embeddings may be less representative of the original context of the text sequence in the input space, and so will exhibit less effective task performance. We note that the divergence in our results here is an artefact of the privacy budget parameter - were we to impose a much more stringent constraint in our MDP results than that in our experimental setup, the embedding deviation would likely be much more in line with that observed in LDP.

5.3. Determining the effect of privacy parameters

In proposing the CAPE model, we proffer the potential for tuning the privacy parameters of the model to account for a range of possible privacy vs. utility scenarios. Since this is also technically possible given the privacy budget parameter ϵ within the LDP model, we also wished to investigate the effect this has.

To determine the effect of adjusting these parameters, we carried out a number of experiments over a range of sensible values drawn from the existing literature: in terms of the weight given to the contribution of the adversarial learning system in the model's overall objective function denoted by the variable λ, we refer to the values established in Ganin et al. (2016), while our range of values for the privacy budget parameter ϵ, which controls the addition of noise under the local differential privacy regime, follows the example of Lyu et al. (2020).

We carry out each run of both models over the English instances from our dataset, using the *bert-base-cased* embedding set. Results for both the basic review score prediction task and simulated attacker task are presented in Table 11.

We determine that base task and attacker performance increases broadly in proportion to the size of the privacy parameter ϵ, as was expected, i.e. laxer privacy budgets provide more limited privacy preservation while exhibiting fewer negative effects on performance. However, we can also see that increasing the weighting of the adversarial training objective by raising the value of λ above 1.0 has significant adverse effects on attacker performance, although this is not true in all cases (see Fig. 9). The behaviour of the adversarial parameter in these experiments is variable—there is no marked trend towards any particular value being preferable across privacy budget settings.

As noted in the literature on Metric Differential Privacy (Fernandes, 2021), the value of the privacy parameter ϵ may be difficult to contextualise outside of the specific metric space in use and cannot be transferred meaningfully across metrics, in contrast to other regimes of differential privacy. In this sense, the effect of the privacy parameter cannot be intuitively understood through comparison to other values and must be established through calibrating experiments against other, better-understood output statistics (Feyisetan et al., 2020). In our case, since we have access to the performance of our attacker process, we use this as a reasonably stable proxy for the privacy effect obtained.

We design an experiment using a subset of 20,000 instances from both our Danish and English datasets, perturbing both the training and test set using the MDP embedding generation process, before testing with our previously-detailed multi-headed

Table 11
Base task and simulated attacker performance as F1-scores for CAPE and LDP models over a range of privacy budgets (ϵ) and adversarial weight parameter values (λ).

CAPE				LDP		
ϵ	λ	Attacker	Task	ϵ	Attacker	Task
0.1	0.1	0.359	0.422	0.1	0.404	0.373
0.1	0.5	0.380	0.379	1	0.438	0.379
0.1	1	0.467	0.382	10	0.438	0.379
0.1	2	0.336	0.380	20	0.438	0.402
1	0.1	0.438	0.417	50	0.476	0.479
1	0.5	0.369	0.410			
1	1	0.427	0.402			
1	2	0.386	0.387			
10	0.1	0.420	0.387			
10	0.5	0.438	0.402			
10	1	0.438	0.387			
10	2	0.438	0.402			
20	0.1	0.420	0.379			
20	0.5	0.420	0.379			
20	1	0.438	0.402			
20	2	0.441	0.379			
50	0.1	0.420	0.387			
50	0.5	0.420	0.379			
50	1	0.420	0.402			
50	2	0.455	0.486			

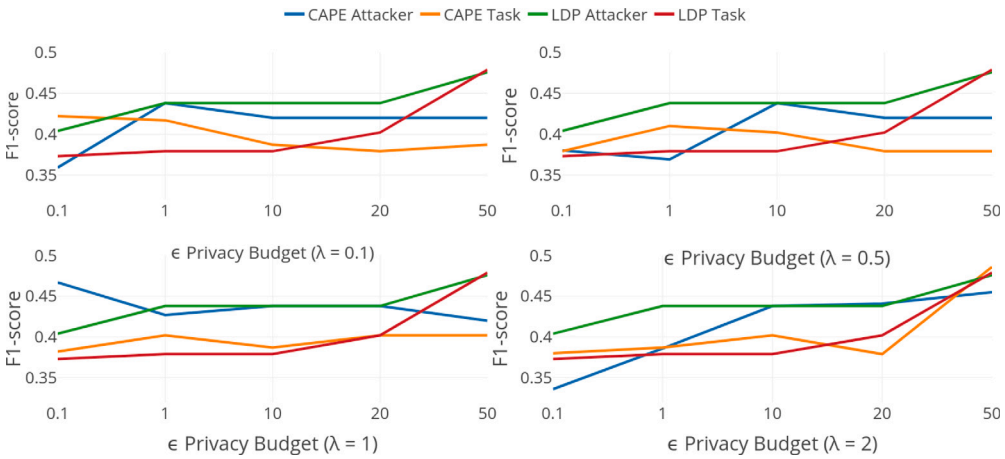

Fig. 9. F1-score for base and simulated attacker tasks for CAPE and LDP models over a range of privacy budget parameters.

Table 12
Privacy outcomes with MDP process across a range of ϵ values.

ϵ value	Danish Attack F1	Danish Task F1	English Attack F1	English Task F1
0.1	0.574	0.376	0.414	0.406
1.0	0.582	0.417	0.419	0.416
10.0	0.668	0.577	0.465	0.440
20.0	0.670	0.578	0.633	0.623
50.0	0.680	0.584	0.604	0.656

classification setup. We obtain results for each demographic attribute with embeds drawn from each PTLM in our experimental set in four cross-verification runs and present a mean average set of results across a range of ϵ values $\in \{0.1, 1.0, 10.0, 20.0, 50.0\}$, as presented in Table 12 and Fig. 10.

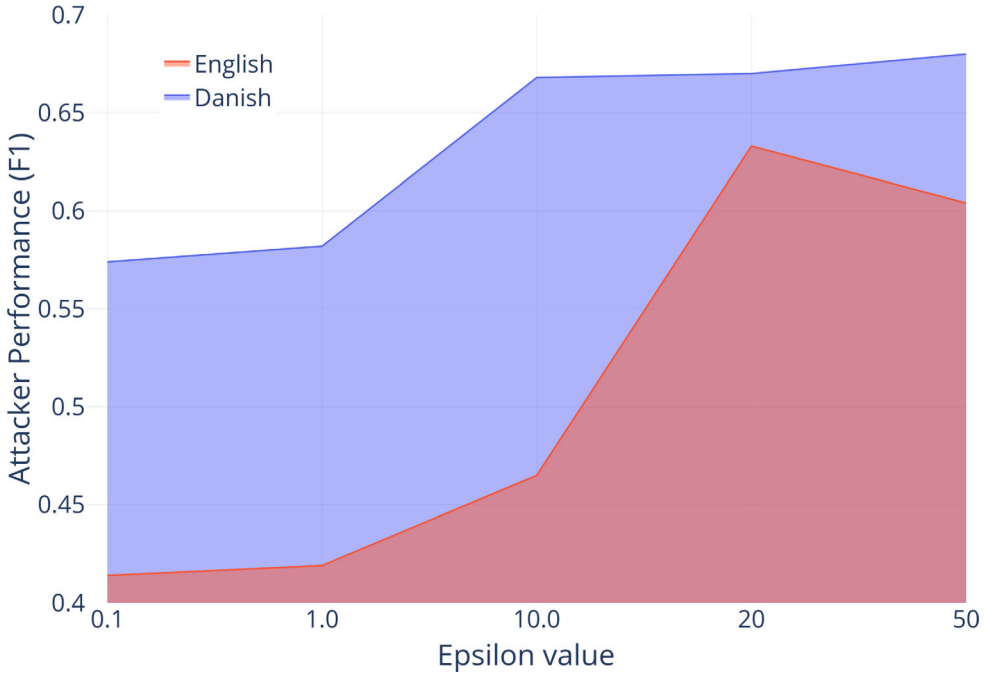

Fig. 10. MDP privacy outcomes across a range of ϵ values for English and Danish.

Table 13
Attacker performance reduction of CAPE strategy across range of classifier model scenarios. Depth refers to the number of layers in the classifier network, width to the number of units in each layer.

Depth	Width	Attacker performance	Depth	Width	Attacker performance
1	50	−42%	4	50	−45%
1	100	−44%	4	100	−44%
1	200	−43%	4	200	−45%
1	500	−45%	4	500	−45%
2	50	−56%	5	50	−44%
2	100	−56%	5	100	−42%
2	200	−44%	5	200	−43%
2	500	−43%	5	500	−44%
3	50	−45%			
3	100	−45%			
3	200	−45%			
3	500	−42%			

5.4. Increasing attack model complexity

Finally, we assess here the impact of increasing classifier complexity and depth on privacy outcomes, both those from non-private encodings and from those privatised using our CAPE model. Since we would expect a deeper classifier network to be prima facie better at distinguishing demographic signal from the noise of everything else encoded in a pre-trained sequence embedding, we would expect a more complex simulated attacker network to therefore exhibit higher information leakage in a non-private setting, and potentially a higher relative drop in attacker performance when privacy preservation is applied.

We carry out a set of experiments across a range of classifier setups for both private and non-private scenarios, using the English language splits from our dataset and obtaining the embeddings from the 'bert-base-cased' language models. We adapt our attacker network by altering the 'depth'—that being the number of layers in the classifier—and also the 'width'—that being the number of units in each layer.

Results are presented in Table 13, as the difference between the F1 score of the attacker classifier when using the CAPE privacy strategy against a non-private baseline across a range of potential classifier setup scenarios. Interestingly, the results do not bear out our assumptions: it is clear that the relative performance of our adversary network does not vary significantly with the complexity of the classifier layers. This may indicate that the attacker network topology is not germane to the detection of the relatively robust

demographic signal from the pre-trained embedding, although this would require additional experimentation with alternate network components to judiciously assess.

6. Conclusion

In this work, we have presented an initial set of results that show an indication of the extent of the information leakage issue for language models. We have indicated the vulnerability of a set of popular models to one aspect of personal privacy erosion, although we note that this only represents a small fraction of the potential malicious use cases presented by this technology. We believe that our title, "You Are What You Write" holds some truth here, since it is clear that the attacker in this case is exploiting the ability of the model to leverage context clues within written text to recover demographic attributes about the author that may pose a greater risk to our privacy than we imagine.

Beyond the previously mentioned findings, we believe we have presented some indicative results for practical solutions that may help ameliorate the issue—a concrete necessity given that many governments are evincing interest in regulating the practice (Aho and Duffield, 2020; Minssen et al., 2020). Not only should this serve to reduce the ability of models to perform this kind of unwanted inference, but it could also enable model owners to guard against other forms of unwanted model behaviour. We have not only proposed a simple framework for assessing the practical consequences of implementing various privacy-preserving mechanisms, but we have also extended that toolset to cover multiple potential areas of interest: alternate languages, language models, and types of private information. We have covered in our language model selection a broad section of popular models which we consider representative (Qiu et al., 2020). Given the speed at which new models are being derived, we would expect this list to change radically in a relatively short space of time.

Based on our experimental evidence, we established the following set of axioms that may provide a useful resource for further research in the field:

- **All pre-trained language models leak demographic data.** We were able to extract useful leaked information from all of the models during our testing and convert that into inferences about user demographics that they could reasonably expect to remain private.
- **The complexity of the language model, and the amount of data used to train it, matters.** More complex language models, and those trained with a more extensive corpus of data, showed marked differences in response to our privacy-preserving methods from smaller peers. The correlation between model size and information leakage is yet to be fully determined and requires further research.
- **Classifier complexity is not strongly linked to attacker performance.** Our tests found no significant correlation between attacker performance and the depth of the classifier network.
- **Local differential privacy significantly impacts model utility.** Applying noise across embeddings at the privacy budget tested necessitated a significant cost to base task performance in our testing, which we can partially mitigate in hybrid models with more relaxed budgets.
- **Metric differential privacy shows a much more moderate impact on performance.** The simple Euclidean metric differential privacy implemented here exhibited much lower utility costs while providing somewhat less impressive privacy results.

As language models continue to become available at greater scale and complexity—during the preparation of this work, the 175 billion parameter GPT-3 model (Brown et al., 2020) was made public—we believe this research, and the objective of achieving measurable and empirically verifiable privacy preservation more broadly, only becomes more necessary. Without a commonly agreed framework for making decisions on the acceptable level of information leakage, we run the risk of undermining user trust.

6.1. Further work

Finally, we would point to the following fruitful lines of inquiry we noted during the preparation of this work. In our experiments, we assessed the privacy performance of differentially private and hybrid models using the same set of private variables as we included in the set seen by all the models at training time. In order to test the general level of privacy provided by DP on unseen variables over adversarial learning, it would be desirable to engineer a scenario where we introduce new variables during evaluation that are not included in the training.

One potential avenue of research we did not pursue relates to the impact of representation dimensionality on information leakage; if it is the case that higher-dimensional encodings capture a more nuanced set of interactions between individual tokens within a sequence, then producing lower-dimensional representations with the same model may provide a shortcut to leaking less additional information. This would require retraining a chosen language model several times. In this work, we also deal with perhaps the most simplistic implementation of metric differential privacy, projecting each embedding in Euclidean space. However, multiple alternate formulations that use different metrics have been proposed (Feyisetan et al., 2019; Xu et al., 2020; Carvalho et al., 2021). Evaluating empirically the distinct outcomes of each distance metric would be a worthy endeavour.

The question of privacy attacks other than the re-identification of personal attributes remains unaddressed by our work. One immediate extension we envisage is to also produce results for membership inference attacks: a form of attack where an adversary uses model outputs to infer whether an individual's data was used in the training set for a model (Shokri et al., 2017; Leino and Fredrikson, 2020).

Table A.14
Language tag and no. of examples in dataset.

Language code	Instances	Language code	Instances
en	1 567 424	oc	907
da	935 629	ro	582
fr	430 103	sl	545
no	184 126	ms	527
de	176 115	id	415
nb	24 767	ht	371
nl	8102	hu	344
sv	7823	tr	330
es	7691	lt	306
it	5891	cy	254
ca	5449	af	244
pl	2469	eu	236
wa	2191	tl	222
pt	1948	zh	192
nn	1946	am	180
eo	1275	br	162
mt	1077	cs	152
fi	1051	ug	121
et	964	sk	110

CRediT authorship contribution statement

Richard Plant: Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Valerio Giuffrida:** Writing – review & editing, Supervision. **Dimitra Gkatzia:** Writing – review & editing, Writing – original draft, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

DG was supported by the EPSRC project 'NLG for low-resource domains' (EP/T024917/1).

Appendix A. Dataset description

Since a long-form description of the Hovy et al. (2015) dataset as proscribed by Bender and Friedman (2018) is not available, we have prepared a description of the characteristics of the dataset to aid in recognising the generalisation potential of our research.

Curation Rationale In order to provide a large-scale textual resource with high coverage of demographic annotations for study (Hovy et al., 2015) crawled the Trustpilot website, extracting publicly available review content as JSON-formatted objects. Data in the released set is restricted to those countries with more than 250,000 users at the time of collection in 2015: Denmark, France, Germany, the United Kingdom, and the United States. Collection was restricted to reviews posted at most seven years prior to the date of collection.

The collected set is augmented in two ways. Firstly, when gender information is not supplied by the reviewer, the researchers attempt to add gender by reference to the existing distribution of first names and genders. If a particular first name appears more than 3 times in the un-augmented set and is correlated with one gender at least 95% of the time, that gender is propagated to all other occurrences of that first name without an attached gender. Secondly, the researchers use the Geonames database to add latitude and longitude information to the free-text location field as reported by the reviewer. In occasions where the canonical location is indeterminate, such as when two towns share the same name, the researchers select the largest town of that name in the selected country.

Language Variety Languages represented within the dataset with more than 100 instances are listed in Table A.14 as represented by ISO639-1 codes. Regional variations are not listed, but for English are likely to include both British English (en-UK) and US English (en-US).

Speaker Demographic Speakers were not directly invited to participate in this set and a full demographic overview is not available. No specific information about the users' income, social or economic status is available explicitly within the dataset.

However, the researchers do have access to the self-reported demographic attributes of age, gender and location. Table A.15 shows the proportion of users in each country for which those attributes are available. Number of instances complete with all available demographic information is broken down per-country in Table A.16 and per-language in Table A.17.

Table A.15
Proportion of users with available demographic attributes (Hovy et al., 2015).

	Users	Age	Gender	Location	All
United Kingdom	1424k	7%	62%	5%	4%
France	741k	3%	53%	2%	1%
Denmark	671k	23%	87%	17%	16%
United States	648k	8%	59%	7%	4%
Netherlands	592k	9%	39%	7%	5%
Germany	329k	8%	47%	6%	4%
Sweden	170k	5%	64%	4%	3%
Italy	132k	10%	61%	8%	6%
Spain	56k	6%	37%	5%	3%
Norway	51k	5%	50%	4%	3%
Belgium	36k	13%	42%	11%	8%
Australia	31k	8%	36%	7%	5%
Finland	16k	6%	36%	5%	3%
Austria	15k	10%	43%	7%	5%
Switzerland	14k	8%	41%	7%	4%
Canada	12k	10%	19%	9%	4%
Ireland	12k	8%	30%	7%	4%

Table A.16
Data instances per country.

Country	Instances
Denmark	467,392
France	24,440
Germany	26,810
United Kingdom	146,381
United States	43,879

Table A.17
Data instances per language.

Language	Train	Test	Val	Total
English	136,226	38,921	19,460	194,607
French	16,170	4620	2310	23,100
German	20,129	5751	2875	28,755
Danish	259,866	74,247	37,123	371,236
Norwegian	51,209	14,630	7315	73,154
Multi-language	35,000	10,000	5000	50,000

Table A.18
Average age per country (Hovy et al., 2015).

	Women			Men		
	Mean	Median	off. median	Mean	Median	off. median
Denmark	38.80	38	41.6	39.07	38	39.8
France	42.03	41	41.2	41.92	41	38.2
Germany	40.64	40	45.3	38.97	38	42.3
United Kingdom	44.51	45	41.5	43.87	43	39.4
United States	40.79	40	38.1	36.70	33	35.5

The mean and median ages for users in the sets for each country are given in Table A.18, in comparison to the official figures for each country as reported in the CIA World Factbook (Central Intelligence Agency, 2021).

Annotator Demographic The dataset is not substantially annotated except where demographic attributes have been inferred using the augmentation techniques described in Appendix A.

Speech Situation All reviews were posted to the Trustpilot website between 2008 and 2015. Messages represent extemporaneous literature describing the experience of users with various businesses, services, products, and other agencies. The public nature of these communications, along with the associated star ratings, indicates that these were regarded by the user as intended for public consumption, with the intended audience presumably prospective users of the target business or service. A secondary target of the speech may also have been the operators of such businesses.

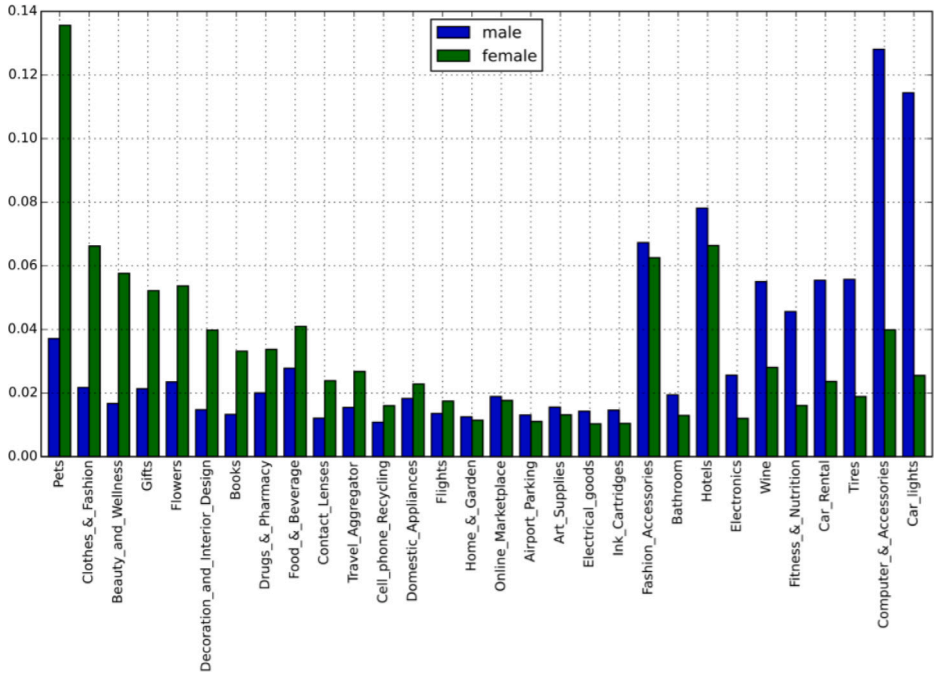

Fig. A.11. Most common review categories by gender (Hovy et al., 2015).

Text Characteristics The majority of texts in this dataset concern reviews of products and retail services such as parking charges or home decoration, rather than other forms of business transaction such as legal or business consultancy. Distribution across the most common categories addressed by the review broken down by gender (Hovy et al., 2015) is illustrated in Fig. A.11.

Appendix B. Language models

B.1. Model details

The language models selected for use in these experiments are listed in Table B.19, grouped by input language. These models were selected based on their popularity as represented by the number of downloads via a popular open-source model-sharing repository.

Pre-training objectives for language models can have a significant impact on downstream behaviour, we have referred to these strategies as follows:

- MLM: Masked Language Modelling, in which a proportion of tokens in the input sequence are obscured, and the model is tasked with predicting the correct token to insert.
- CLM: Casual Language Modelling, in which the model is given a sequence of tokens as input and has to predict the next token in the output sequence.
- TLM: Translation Language Modelling, in which firstly aligned sequences in different languages are concatenated, then the masked token prediction task is performed on the full input sequence.
- NSP: Next Sentence Prediction, in which two (potentially unrelated) sequences are concatenated and the model predicts whether one sequence follows the other.
- SOP: Sentence Order Prediction, in which the model is presented with two sequences, and must predict the correct ordering.
- MKD: Multilingual Knowledge Distillation, in which a teacher model translates an original sequence into an embedding, while a student model learns to map translations of the original input onto the same embedding space.

B.2. Hyper-parameter selection

Hyper-parameter selection was carried out for the classification model based on the initial task performance (sentiment prediction) in the Baseline scenario, that is without the introduction of privacy-preserving interventions. Optimal parameters were selected for increased task performance in the metrics of interest as detailed earlier in the paper. Details on the parameter, selected experimental values, and the set of parameters from which they were selected if any, are contained in Table B.20.

Some parameters, such as the maximum input sequence, vocabulary size, fully-connected (dense) layer size, activation function, loss function, and optimiser parameters were set for the classifier based on the values suggested by Beigi et al. (2019) in their initial

Table B.19
Language models used.

Language	Model	Parameters	Training corpus	Objective	Embedding dimension
English	bert-base-cased	110M	16 GB	MLM & NSP	768
	bert-base-multilingual-cased	110M	Unknown	MLM & NSP	768
	roberta-base	125M	160 GB	MLM	768
	albert-base	11M	160 GB	MLM & SOP	128
	gpt2	124M	40 GB	CLM	768
French	flaubert_base_cased	138M	71 GB	MLM	768
	bert-base-multilingual-cased	110M	Unknown	MLM & NSP	768
	camembert-base	110M	138 GB	MLM	768
	belgpt2	124M	60 GB	CLM	768
German	bert-base-german-cased	110M	12 GB	MLM	768
	bert-base-multilingual-cased	110M	Unknown	MLM & NSP	768
	cross-en-de-roberta-sentence-transformer	250M	Unknown	TLM	768
	german-gpt2	124M	16 GB	CLM	768
Danish	danish-bert-botxo	110M	9.5 GB	MLM	768
	bert-base-multilingual-cased	110M	Unknown	MLM & NSP	768
	roberta-base-danish	125M	107 GB	MLM	768
Norwegian	norbert	110M	2B tokens	MLM	768
	bert-base-multilingual-cased	110M	Unknown	MLM & NSP	768
	nordic-roberta-wiki	125M	16 GB	MLM	768
Multilingual	bert-base-multilingual-cased	110M	Unknown	MLM & NSP	768
	paraphrase-xlm-r-multilingual-v1	250M	Unknown	TLM	768
	distiluse-base-multilingual-cased-v2	50M	Unknown	MKD	3072

Table B.20
Model hyper-parameters.

Parameter	Value	Selection set
Random seed	42	
Max. input sequence length (tokens)	512	
Max. vocabulary size (tokens)	10,000	
Training batch size	32	[16, 32, 64, 128]
Testing batch size	64	[16, 32, 64, 128]
Max. training length (epochs)	100	
Cross-validation requirement (runs)	4	
Classifier hidden layer size (units)	64	[32, 64, 128, 256]
Classifier dense layer size (units)	200	
Classifier activation function	Softmax	
Classifier loss function	Categorical cross-entropy	
Loss label smoothing	0	
Classifier optimizer	Adam	
Optimizer learning rate	0.001	[0.1, 0.01, 0.001, 0.0001]
Optimizer decay rate (1st moment)	0.9	
Optimizer decay rate (2nd moment)	0.999	
Optimizer stability constant (ϵ)	1e−07	

study, which we adopt as a stable basis for comparison in our work. Other parameters such as batch size were dictated by the execution environment and available resources.

Appendix C. Full results tables

See Tables C.21–C.27.

Table C.21
English results. Gl: GloVe, Ko: Komninos, Be: BERT, MB: Multilingual BERT, Ro: RoBERTa, Al: ALBERT, GPT: GPT-2.

Model	Gender	Location	Age	σ^2	Model	Gender	Location	Age	σ^2
Baseline					**LDP**				
Gl	0.674	0.661	0.674	0.021	Gl	0.374	0.649	0.449	0.014
Ko	0.668	0.630	0.680	0.024	Ko	0.374	0.649	0.448	0.014
Be	0.737	0.749	0.746	0.027	Be	0.375	**0.649**	0.449	0.020
MB	0.712	0.696	0.691	0.021	MB	0.375	0.709	0.454	0.030
Ro	0.765	0.771	0.773	0.026	Ro	0.375	0.649	0.449	0.020
Al	0.779	0.765	0.754	0.029	Al	0.646	0.819	0.457	0.033
GPT	0.658	0.730	0.753	0.024	GPT	0.561	0.708	0.449	0.017
CGT					**MDP**				
Gl	0.375	0.649	0.449	0.020	Gl	0.572	0.719	0.456	0.017
Ko	0.375	0.649	0.449	0.020	Ko	0.572	0.703	0.459	0.015
Be	0.375	0.649	0.449	0.020	Be	0.653	0.798	0.498	0.023
MB	0.375	0.649	0.449	0.020	MB	0.635	0.738	0.472	0.018
Ro	0.375	0.649	0.449	0.020	Ro	0.664	0.809	0.501	0.024
Al	0.375	0.649	0.449	0.020	Al	0.645	0.790	0.479	0.024
GPT	0.375	0.649	0.449	0.020	GPT	0.646	0.768	0.491	0.019
GR					**CAPE**				
Gl	0.375	0.649	0.449	0.020	Gl	0.375	0.649	0.449	0.013
Ko	0.378	0.649	0.449	0.020	Ko	0.375	0.649	0.052	0.060
Be	0.375	0.649	0.448	0.020	Be	0.507	0.649	0.339	0.024
MB	0.377	0.649	0.448	0.020	MB	0.378	0.649	**0.278**	0.082
Ro	0.375	0.649	0.449	0.020	Ro	**0.350**	0.649	0.449	0.023
Al	0.374	0.649	0.449	0.020	Al	0.356	0.649	0.424	0.017
GPT	0.376	0.648	0.448	0.020	GPT	0.375	0.649	0.449	0.020

ANOVA metrics	SS	df	F	P	F-crit
Row	4.024	41	18.674	$1.334e^{-27}$	1.537
Column	1.331	2	126.587	$8.517e^{-26}$	3.108
Total	5.786	125			

Table C.22
French results. GPT: GPT-2, MB, Multilingual BERT, CB: CamemBERT, FB: FlauBERT.

Model	Gender	Location	Age	σ^2	Model	Gender	Location	Age	σ^2
Baseline					**LDP**				
FB	0.680	0.985	0.414	0.082	FB	0.405	0.985	0.419	0.110
MB	0.658	0.985	0.405	0.085	MB	0.404	0.985	0.419	0.110
CB	0.724	0.985	0.447	0.072	CB	0.404	0.985	0.419	0.110
GPT	0.645	0.985	0.376	0.093	GPT	0.636	0.985	0.419	0.082
CGT					**MDP**				
FB	0.404	0.985	0.419	0.110	FB	0.571	0.985	0.417	0.086
MB	0.404	0.985	0.419	0.110	MB	0.624	0.985	0.419	0.082
CB	0.404	0.985	0.419	0.110	CB	0.712	0.985	0.443	0.073
GPT	0.404	0.985	0.419	0.110	GPT	0.659	0.985	0.414	0.082
GR					**CAPE**				
FB	0.410	0.985	0.414	0.109	FB	0.499	0.985	**0.383**	0.102
MB	0.404	0.985	0.419	0.110	MB	0.404	0.985	0.404	0.113
CB	0.404	0.985	0.419	0.110	CB	**0.403**	0.985	0.414	0.111
GPT	0.405	**0.984**	0.405	0.112	GPT	0.435	0.985	0.419	0.104

ANOVA metrics	SS	df	F	P	F-crit
Row	0.125	23	1.047	0.434	1.767
Column	4.514	2	435.333	$1.3e^{-30}$	3.200
Total	4.878	71			

Table C.23
German results. Be: BERT, MB: Multilingual BERT, Ro: RoBERTa, GPT: GPT-2.

Model	Gender	Location	Age	σ^2	Model	Gender	Location	Age	σ^2
Baseline					**LDP**				
Be	0.683	0.985	0.391	0.088	Be	0.528	0.985	0.328	0.113
MB	0.614	0.985	0.413	0.084	MB	0.528	0.985	0.328	0.113
Ro	0.652	0.985	0.435	0.079	Ro	0.528	0.985	0.328	0.113
GPT	0.677	0.985	0.449	0.072	GPT	0.528	0.985	0.328	0.113
CGT					**MDP**				
Be	0.528	0.985	0.328	0.113	Be	0.673	0.985	0.401	0.086
MB	0.528	0.985	0.328	0.113	MB	0.632	0.985	0.421	0.081
Ro	0.528	0.985	0.328	0.113	Ro	0.640	0.985	0.440	0.076
GPT	0.528	0.985	0.328	0.113	GPT	0.680	0.985	0.424	0.079
GR					**CAPE**				
Be	0.528	0.985	**0.146**	0.177	Be	0.528	0.985	0.328	0.114
MB	0.529	0.985	0.249	0.138	MB	0.528	0.985	0.328	0.113
Ro	0.544	0.984	0.317	0.115	Ro	**0.527**	0.985	0.328	0.114
GPT	0.528	0.985	0.328	0.113	GPT	0.528	0.985	0.198	0.156

ANOVA metrics	SS	df	F	P	F-crit
Row	0.121	23	2.717	0.002	1.767
Column	5.098	2	1314.380	$2.61e^{-41}$	3.200
Total	5.308	71			

Table C.24
Danish results. Be: BERT, MB: Multilingual BERT, Ro: RoBERTa.

Model	Gender	Location	Age	σ^2	Model	Gender	Location	Age	σ^2
Baseline					**LDP**				
Be	0.689	0.996	0.392	0.091	Be	0.608	0.996	0.341	0.108
MB	0.622	0.996	0.334	0.110	MB	0.430	0.996	0.225	0.159
Ro	0.784	0.996	0.300	0.122	Ro	0.430	0.996	0.257	0.149
CGT					**MDP**				
Be	0.430	0.996	0.272	0.145	Be	0.692	0.996	0.415	0.084
MB	0.430	0.996	0.272	0.145	MB	0.649	0.996	0.337	0.109
Ro	0.430	0.996	0.272	0.145	Ro	0.535	0.996	0.260	0.138
GR					**CAPE**				
Be	0.430	0.996	**0.010**	0.249	Be	0.430	0.996	0.121	0.197
MB	0.430	0.996	0.262	0.148	MB	**0.430**	0.996	0.197	0.169
Ro	0.430	0.996	0.098	0.206	Ro	0.430	**0.995**	0.272	0.145

ANOVA metrics	SS	df	F	P	F-crit
Row	0.208	17	2.688	0.007	1.933
Column	5.082	2	559.094	$9.755e^{-27}$	3.276
Total	5.444	53			

Table C.25
Norwegian results. Be: BERT, MB: Multilingual BERT, Ro: RoBERTa.

Model	Gender	Location	Age	σ^2	Model	Gender	Location	Age	σ^2
Baseline					**LDP**				
Be	0.590	0.995	0.308	0.119	Be	0.407	0.995	0.249	0.154
MB	0.589	0.995	0.310	0.118	MB	0.407	0.995	0.206	0.168
Ro	0.620	0.995	0.301	0.120	Ro	0.407	0.995	0.212	0.166
CGT					**MDP**				
Be	0.407	0.995	0.254	0.153	Be	0.606	0.995	0.285	0.126
MB	0.407	0.995	0.254	0.153	MB	0.571	0.995	0.279	0.129
Ro	0.407	0.995	0.254	0.153	Ro	0.612	0.995	0.276	0.129
GR					**CAPE**				
Be	0.407	0.995	0.230	0.160	Be	**0.326**	0.995	0.252	0.167
MB	0.407	**0.994**	0.152	0.187	MB	0.453	0.995	0.203	0.164
Ro	0.407	0.995	**0.087**	0.212	Ro	0.509	0.995	0.156	0.199

ANOVA metrics	SS	df	F	P	F-crit
Row	0.127	17	2.211	0.024	1.933
Column	5.443	2	803.777	$2.376e^{-29}$	3.276
Total	5.685	53			

Table C.26
Multi-language results. MB: Multilingual BERT, XLM: Multilingual RoBERTa, USE: Multilingual Universal Sentence Encoder.

Model	Gender	Location	Age	σ^2	Model	Gender	Location	Age	σ^2
Baseline					**LDP**				
MB	0.589	0.598	0.589	$2.665e^{-5}$	MB	0.412	0.860	0.336	0.080
USE	0.617	0.612	0.608	$1.847e^{-5}$	USE	0.412	0.439	0.336	0.003
XLM	0.640	0.638	0.643	$5.078e^{-6}$	XLM	0.420	0.718	0.325	0.042
CGT									
MB	0.412	0.859	0.336	0.080					
USE	0.530	0.768	0.336	0.047					
XLM	0.574	0.749	0.336	0.043					
GR					**CAPE**				
MB	0.412	0.576	0.336	0.015	MB	0.412	0.442	0.336	0.075
USE	0.412	**0.439**	0.335	0.003	USE	**0.412**	0.439	0.316	0.004
XLM	0.412	0.440	0.320	0.004	XLM	0.460	0.439	**0.266**	0.011

ANOVA metrics	SS	df	F	P	F-crit
Row	0.333	14	1.863	0.078	2.064
Column	0.456	2	17.872	$9.957e^{-6}$	3.340
Total	1.147	44			

Table C.27
Base task performance F1-score per model per language, averaged across all runs.

Model	Base task F1-score	Model	Base task F1-score
English		**Danish**	
AE	0.530	FX	0.534
FX	0.714	CGT	0.317
CGT	0.491	GR	0.516
GR	0.705	LDP	0.416
LDP	0.600	MDP	0.529
MDP	0.648	CAPE	0.368
CAPE	0.493	**Norwegian**	
French		FX	0.562
FX	0.574	CGT	0.424
CGT	0.362	GR	0.554
GR	0.564	LDP	0.441
LDP	0.411	MDP	0.551
MDP	0.575	CAPE	0.424
CAPE	0.369	**Multilingual**	
German		FX	0.615
FX	0.613	CGT	0.484
CGT	0.450	GR	0.590
GR	0.599	LDP	0.440
LDP	0.468	CAPE	0.416
MDP	0.612		
CAPE	0.416		

Data availability

Data will be made available on request.

References

Abdaoui, A., Pradel, C., Sigel, G., 2020. Load what you need: Smaller versions of mutililingual BERT. In: Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing. Association for Computational Linguistics, Online, pp. 119–123. http://dx.doi.org/10.18653/v1/2020.sustainlp-1.16.

Aho, B., Duffield, R., 2020. Beyond surveillance capitalism: Privacy, regulation and big data in Europe and China. Econ. Soc. 49 (2), 187–212, URL: https://doi.org/10.1080/03085147.2019.1690275. doi:10/gjhf78. Publisher: Routledge _eprint: https://doi.org/10.1080/03085147.2019.1690275.

Aletras, N., Chamberlain, B.P., 2018. Predicting twitter user socioeconomic attributes with network and language information. In: HT 2018 - Proceedings of the 29th ACM Conference on Hypertext and Social Media. Association for Computing Machinery, Inc, pp. 20–24. http://dx.doi.org/10.1145/3209542.3209577, URL: http://arxiv.org/abs/1804.04095. arXiv:1804.04095.

Alnasser, W., Beigi, G., Liu, H., 2021. Privacy preserving text representation learning using BERT. In: Thomson, R., Hussain, M.N., Dancy, C., Pyke, A. (Eds.), Social, Cultural, and Behavioral Modeling. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 91–100, doi:10/gmmzv3. URL:10/gmmzv3.

Anwar, M., 2021. Supporting privacy, trust, and personalization in online learning. Int. J. Artif. Intell. Educ. 31 (4), 769–783, doi:10/gnrdp8. URL:.

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D., 2019. Invariant risk minimization. arXiv preprint arXiv:1907.02893.

Arora, A., Arora, A., 2022. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthc. J. 9 (2), 190–193. http://dx.doi.org/10.7861/fhj.2022-0013, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9345230/.

Bae, H., Jung, D., Choi, H.S., Yoon, S., 2020. AnomiGAN: Generative adversarial networks for anonymizing private medical data. In: Pacific Symposium on Biocomputing. Vol. 25, World Scientific Publishing Co. Pte Ltd, pp. 563–574. http://dx.doi.org/10.1142/9789811215636_0050, URL: http://arxiv.org/abs/1901.11313. arXiv:1901.11313 Issue: 2020 ISSN: 23356936.

Beigi, G., Shu, K., Guo, R., Wang, S., Liu, H., 2019. Privacy preserving text representation learning. In: HT 2019 - Proceedings of the 30th ACM Conference on Hypertext and Social Media. pp. 275–276. http://dx.doi.org/10.1145/3342220.3344925, URL: http://arxiv.org/abs/1907.03189. arXiv:1907.03189.

Bender, E.M., Friedman, B., 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Trans. Assoc. Comput. Linguist. 6, 587–604, doi:10/gft5d7. URL: https://www.aclweb.org/anthology/Q18-1041.

Benjamin, M., 2018. Hard numbers: Language exclusion in computational linguistics and natural language processing. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC 2018, European Language Resources Association (ELRA), Paris, France, p. 6.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: Advances in Neural Information Processing Systems. Vol. 33, Curran Associates, Inc., pp. 1877–1901, URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cai, Z., Xiong, Z., Xu, H., Wang, P., Li, W., Pan, Y., 2021. Generative adversarial networks: A survey toward private and secure applications. ACM Comput. Surv. 54 (6), 132:1–132:38. http://dx.doi.org/10.1145/3459992.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D., 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Security Symposium. USENIX Association, pp. 267–284, URL: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini. arXiv:1802.08232.

Carvalho, R.S., Vasiloudis, T., Feyisetan, O., 2021. TEM: High utility metric differential privacy on text. arXiv:2107.07928 [cs]. URL: http://arxiv.org/abs/2107.07928.

Central Intelligence Agency, 2021. Median age - the world factbook. URL: https://www.cia.gov/the-world-factbook/field/median-age/,

Chan, B., Schweter, S., Möller, T., 2020. German's next language model. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 6788–6796, doi:10/gnh2t8. URL:10/gnh2t8.

Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C., 2013. Broadening the scope of differential privacy using metrics. In: PETS 2013: Privacy Enhancing Technologies. In: Lecture Notes in Computer Science, vol. 7981, Springer, pp. 82–102. http://dx.doi.org/10.1007/978-3-642-39077-7_5, URL: https://hal.inria.fr/hal-00767210.

Chatzikokolakis, K., Palamidessi, C., Stronati, M., 2015. Constructing elastic distinguishability metrics for location privacy. Proc. Priv. Enhanc. Technol. 2015 (2), 156–170. http://dx.doi.org/10.1515/popets-2015-0023, Publisher: Sciendo Section: Proceedings on Privacy Enhancing Technologies.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N., 2022. PaLM: Scaling language modeling with pathways. http://dx.doi.org/10.48550/arXiv.2204.02311, URL: http://arxiv.org/abs/2204.02311. arXiv:2204.02311 [cs].

Cimino, A., Dell'Orletta, F., Nissim, M., 2020. TAG-it@ EVALITA 2020: Overview of the topic, age, and gender prediction task for Italian: Evaluation campaign of natural language processing and speech tools for Italian. In: Basile, V., Croce, D., Di Maro, M., Passaro, L.C. (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop. EVALITA 2020, CEUR Workshop Proceedings (CEUR-WS.org).

Coavoux, M., Narayan, S., Cohen, S.B., 2018. Privacy-preserving neural representations of text. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 1–10. http://dx.doi.org/10.18653/v1/D18-1001, URL: https://aclanthology.org/D18-1001.

Colleoni, E., Rozza, A., Arvidsson, A., 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. J. Commun. 64 (2), 317–332, doi:10/f5xvkm. URL:10/f5xvkm.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., Wang, T., 2018. Privacy at scale: Local differential privacy in practice. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery, New York, New York, USA, pp. 1655–1658. http://dx.doi.org/10.1145/3183713.3197390, URL: http://dl.acm.org/citation.cfm?doid=3183713.3197390. ISSN: 07308078.

Cornips, L., Gregersen, F., 2017. Comparative studies of variation in the use of grammatical gender in the Danish and Dutch DP in the speech of youngsters: Free versus bound morphemes. Cross-Linguist. Influ. Biling. 101–126, URL: https://www.jbe-platform.com/content/books/9789027265616-sibil.52.06cor. Publisher: John Benjamins.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. 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, pp. 4171–4186, doi:10/ggbwf6. URL:10/ggbwf6.

Dhingra, B., Shallue, C.J., Norouzi, M., Dai, A.M., Dahl, G.E., 2018. Embedding text in hyperbolic spaces. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Student Research Workshop. Association for Computational Linguistics (ACL), pp. 59–69. http://dx.doi.org/10.18653/v1/w18-1708, URL: http://arxiv.org/abs/1806.04313. arXiv:1806.04313.

Doi, S., Mizuno, T., Fujiwara, N., 2020. Estimation of socioeconomic attributes from location information. J. Comput. Soc. Sci. http://dx.doi.org/10.1007/s42001-020-00073-w, Publisher: Springer Science and Business Media LLC.

Dwork, C., 2006. Differential privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II. ICALP '06, Springer-Verlag, Berlin, Heidelberg, pp. 1–12. http://dx.doi.org/10.1007/11787006_1.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Calibrating noise to sensitivity in private data analysis. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, vol. 3876, pp. 265–284. http://dx.doi.org/10.1007/11681878_14.

Elazar, Y., Goldberg, Y., 2018. Adversarial removal of demographic attributes from text data. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 11–21, URL: http://arxiv.org/abs/1808.06640. arXiv:1808.06640.

Emmery, C., Kádár, Á., Chrupała, G., 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp. 2388–2402, URL: https://aclanthology.org/2021.eacl-main.203.

Erlingsson, Ú., Pihur, V., Korolova, A., 2014. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the ACM Conference on Computer and Communications Security. pp. 1054–1067. http://dx.doi.org/10.1145/2660267.2660348, arXiv:1407.6981 ISSN: 15437221.

Fernandes, N., 2021. Differential Privacy for Metric Spaces: Information-Theoretic Models for Privacy and Utility with New Applications to Metric Domains (Ph.D. thesis). Institut Polytechnique de Paris; Macquarie University, Sydney, Australia, URL: https://tel.archives-ouvertes.fr/tel-03344453.

Fernandes, N., Dras, M., McIver, A., 2019. Generalised differential privacy for text document processing. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, vol. 11426, Springer Verlag, pp. 123–148. http://dx.doi.org/10.1007/978-3-030-17138-4_6, arXiv:1811.10256 ISSN: 16113349.

Feyisetan, O., Balle, B., Drake, T., Diethe, T., 2020. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In: WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining. Association for Computing Machinery, Inc, pp. 178–186. http://dx.doi.org/10.1145/3336191.3371856, URL: https://dl.acm.org/doi/10.1145/3336191.3371856. arXiv:1910.08902.

Feyisetan, O., Diethe, T., Drake, T., 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In: Proceedings - IEEE International Conference on Data Mining. ICDM, IEEE Computer Society, pp. 210–219, doi:10/gmpdj8. URL: https://www.computer.org/csdl/proceedings-article/icdm/2019/460400a210/1h5XJ1RyTSM.

Frankowski, D., Cosley, D., Sen, S., Terveen, L., Riedl, J., 2006. You are what you say: privacy risks of public mentions. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06, Association for Computing Machinery, New York, NY, USA, pp. 565–572, doi:10/dq65cj. URL: https://doi.org/10.1145/1148170.1148267.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (1), 2096–2130.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6572.

Gordon, J., Rawlinson, D., Ahmad, S., 2020. Long distance relationships without time travel: Boosting the performance of a sparse predictive autoencoder in sequence modeling. In: Schilling, F.-P., Stadelmann, T. (Eds.), Artificial Neural Networks in Pattern Recognition. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 52–64, doi:10/gmr4zj. URL:10/gmr4zj.

Guo, H., Dolhansky, B., Hsin, E., Dinh, P., Wang, S., Ferrer, C.C., 2019. Deep poisoning functions: Towards robust privacy-safe image data sharing. arXiv preprint URL: http://arxiv.org/abs/1912.06895. arXiv:1912.06895.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., Yuan, J., Zhao, W.X., Zhu, J., 2021. Pre-trained models: Past, present and future. AI Open doi:10/gnzb2d. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000231.

Henriksen-Bulmer, J., Jeary, S., 2016. Re-identification attacks—A systematic literature review. Int. J. Inf. Manage. 36 (6, Part B), 1184–1192. http://dx.doi.org/10.1016/j.ijinfomgt.2016.08.002, URL: https://www.sciencedirect.com/science/article/pii/S0268401215301262.

Hord, L., 2016. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. West. Pap. Linguist./Cah. Linguist. West. 3 (1), URL: https://ir.lib.uwo.ca/wpl_clw/vol3/iss1/4.

Horvitz, E., Mulligan, D., 2015. Data, privacy, and the greater good. Science 349 (6245), 253–255, doi:10/f7kcb3. URL: Publisher: American Association for the Advancement of Science.

Hovy, D., Johannsen, A., Søgaard, A., 2015. User review sites as a resource for large-scale sociolinguistic studies. In: WWW '15: Proceedings of the 24th International Conference on World Wide Web. pp. 452–461. http://dx.doi.org/10.1145/2736277.2741141.

Iliev, G., Genov, A., 2012. Expanding parallel resources for medium-density languages for free. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation. LREC'12, Istanbul, Turkey, p. 7.

Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P., 2014. Political ideology detection using recursive neural networks. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp. 1113–1122, doi:10/ggm374. URL:10/ggm374.

Jaiswal, M., Provost, E.M., 2020. Privacy enhanced multimodal neural representations for emotion recognition. Proc. AAAI Conf. Artif. Intell. 34 (05), 7985–7993, doi:10/gj47dx. URL:10/gj47dx. Number: 05.

Jorgensen, Z., Yu, T., Cormode, G., 2015. Conservative or liberal? Personalized differential privacy. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 1023–1034. http://dx.doi.org/10.1109/ICDE.2015.7113353, ISSN: 2375-026X.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M., 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 6282–6293. http://dx.doi.org/10.18653/v1/2020.acl-main.560, URL: http://arxiv.org/abs/2004.09095. arXiv:2004.09095 Publication Title: arXiv.

Kairouz, P., Bonawitz, K., Ramage, D., 2016. Discrete distribution estimation under local privacy. In: 33rd International Conference on Machine Learning. ICML 2016, Vol. 5, pp. 3607–3633, arXiv:1602.07387.

Kaneko, M., Bollegala, D., 2019. Gender-preserving debiasing for pre-trained word embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 1641–1650, doi:10/ggzshs. URL:10/ggzshs.

Kokovidis, A., 2015. Gendered Discourse in German Chatroom Conversations: the Use of Modal Particles by Young Adults (Ph.D. thesis). Boston University, Boston, MA, URL: https://open.bu.edu/handle/2144/14000. Accepted: 2016-01-13T18:28:31Z.

Komninos, A., Manandhar, S., 2016. Dependency based embeddings for sentence classification tasks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp. 1490–1500, doi:10/ggwcv8. URL:10/ggwcv8.

Kotze, H., Janssen, B., Koolen, C., Plas, L.v.d., Egdom, G.-W.v., 2021. Norms, affect and evaluation in the reception of literary translations in multilingual online reading communities: Deriving cognitive-evaluative templates from big data. Transl. Cognit. Behav. 4 (2), 147–186. http://dx.doi.org/10.1075/tcb.00060.kot, URL: https://www.jbe-platform.com/content/journals/10.1075/tcb.00060.kot. Publisher: John Benjamins.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., Herzog, S.M., 2021. Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, Great Britain, and the United States. Hum. Soc. Sci. Commun. 8 (1), 1–11, doi:10/gmgpfd. URL:10/gmgpfd.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2019. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. Technical Report, Google Research, URL: https://arxiv.org/abs/1909.11942v6.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D., 2020. FlauBERT: Unsupervised language model pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 2479–2490, URL: https://aclanthology.org/2020.lrec-1.302.

Leino, K., Fredrikson, M., 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In: Proceedings of the 29th USENIX Security Symposium. USENIX Association, pp. 1605–1622, URL: https://www.usenix.org/conference/usenixsecurity20/presentation/leino. arXiv:1906.11798.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880, URL: https://virtual.acl2020.org/paper_main.703.html.

Li, Y., Baldwin, T., Cohn, T., 2018. Towards robust and privacy-preserving text representations. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). Vol. 2, Association for Computational Linguistics (ACL), pp. 25–30. http://dx.doi.org/10.18653/v1/p18-2005, arXiv:1805.06093.

Li, T., Li, N., 2009. On the tradeoff between privacy and utility in data publishing. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 517–525. http://dx.doi.org/10.1145/1557019.1557079.

Li, R., Li, X., Chen, G., Lin, C., 2020b. Improving variational autoencoder for text modelling with timestep-wise regularisation. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 2381–2397, doi:10/gmr4zh. URL:10/gmr4zh.

Li, H., Tu, M., Huang, J., Narayanan, S., Georgiou, P., 2020a. Speaker-invariant affective representation learning via adversarial training. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7144–7148, doi:10/ghjxjt. URL:10/ghjxjt. ISSN: 2379-190X.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs]. URL: http://arxiv.org/abs/1907.11692.

Lohndal, T., Westergaard, M., 2021. Grammatical gender: Acquisition, attrition, and change. J. Ger. Linguist. 33 (1), 95–121, doi:10/gnh7cq. URL: https://www.cambridge.org/core/journals/journal-of-germanic-linguistics/article/grammatical-gender-acquisition-attrition-and-change/BC65E747C87CE7B00ED50C472332C1BB. Publisher: Cambridge University Press.

Lyu, L., He, X., Li, Y., 2020. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2355–2365. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.213, URL: https://aclanthology.org/2020.findings-emnlp.213/. arXiv:2010.01285.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (86), 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Madras, D., Creager, E., Pitassi, T., Zemel, R., 2018. Learning adversarially fair and transferable representations. In: Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80, PMLR, pp. 3384–3393, URL: http://arxiv.org/abs/1802.06309. arXiv:1802.06309.

Mahawaga Arachchige, P.C., Bertok, P., Khalil, I., Liu, D., Camtepe, S., Atiquzzaman, M., 2020. Local differential privacy for deep learning. IEEE Internet Things J. 7 (7), 5827–5842. http://dx.doi.org/10.1109/JIOT.2019.2952146, arXiv:1908.02997.

Maheshwari, G., Denis, P., Keller, M., Bellet, A., 2022. Fair NLP Models with Differentially Private Text Encoders. Technical Report, arXiv:2205.06135. arXiv. URL: http://arxiv.org/abs/2205.06135. arXiv:2205.06135 [cs] type: article.

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B., 2020. CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 7203–7219, doi:10/ghpch6. URL:10/ghpch6.

Minssen, T., Gerke, S., Aboy, M., Price, N., Cohen, G., 2020. Regulatory responses to medical machine learning. J. Law Biosci. 7 (1), lsaa002, doi:10/gjc94d. URL:.

Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: Proceedings - IEEE Symposium on Security and Privacy. pp. 111–125. http://dx.doi.org/10.1109/SP.2008.33, ISSN: 10816011.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Catanzaro, B., 2021. Scaling Language Model Training to a Trillion Parameters Using Megatron. Technical Report, Nvidia, URL: https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S.O., Muhammad, S.H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R.A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L.J., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J.T., Abbott, J., Orife, I., Ezeani, I., Dangana, I.A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Murhabazi, E., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C., Dossou, B., Sibanda, B., Bassey, B.I., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., Bashir, A., 2020. Participatory research for low-resourced machine translation: A case study in african languages. arXiv:2010.02353 [cs]. URL: http://arxiv.org/abs/2010.02353.

Ortiz Suárez, P.J., Romary, L., Sagot, B., 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 1703–1714, doi:10/gm5dsn. URL:10/gm5dsn.

Ortiz Suárez, P.J., Sagot, B.t., Romary, L., 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In: Bański, P., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., Kupietz, M., Lüngen, H., Iliadi, C. (Eds.), Proceedings of the Workshop on Challenges in the Management of Large Corpora 2019. In: Proceedings of the workshop on challenges in the management of large corpora (CMLC-7) 2019. Cardiff, 22nd july 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, pp. 9–16, doi:10/gjs47x. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1532–1543. http://dx.doi.org/10.3115/v1/D14-1162, URL: http://aclweb.org/anthology/D14-1162.

Peters, M.E., Ruder, S., Smith, N.A., 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In: Proceedings of the 4th Workshop on Representation Learning for NLP. RepL4NLP-2019, Association for Computational Linguistics, Florence, Italy, pp. 7–14, doi:10/ggvxqn. URL:10/ggvxqn.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A., 2019. Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, pp. 2463–2473, doi:10/ggvxn8. URL: https://aclanthology.org/D19-1250.

Phan, N.H., Thai, M.T., Jin, R., Hu, H., Dou, D., 2019. Scalable differential privacy with certified robustness in adversarial learning. In: Proceedings of the 37th International Conference on Machine Learning. Vol. 119, PMLR, pp. 7683–7694, URL: https://arxiv.org/abs/1903.09822. Publication Title: arXiv.

Pires, T., Schlinger, E., Garrette, D., 2019. How multilingual is multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 4996–5001, doi:10/ghm9kf. URL:10/ghm9kf.

Plant, R., Gkatzia, D., Giuffrida, V., 2021. CAPE: Context-aware private embeddings for private language learning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 7970–7978, URL: https://aclanthology.org/2021.emnlp-main.628.

Prabhumoye, S., Boldt, B., Salakhutdinov, R., Black, A.W., 2021. Case study: Deontological ethics in NLP. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp. 3784–3798, doi:10/gntp5m. URL: https://aclanthology.org/2021.naacl-main.297.

Přibáň, P., Steinberger, J., 2021. Are the multilingual models better? Improving czech sentiment with transformers. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. RANLP 2021, INCOMA Ltd., Held Online, pp. 1138–1149, URL: https://aclanthology.org/2021.ranlp-main.128.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., Neubig, G., 2018. When and why are pre-trained word embeddings useful for neural machine translation? In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 529–535, doi:10/gf7fc8. URL:10/gf7fc8.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X., 2020. Pre-trained models for natural language processing: A survey. Sci. China Technol. Sci. 63 (10), 1872–1897, doi:10/ghqkn2. URL:10/ghqkn2.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2018. Language models are unsupervised multitask learners.

Rao, J., Rao, J., Rohatgi, P., 2000. Can pseudonymity really guarantee privacy? In: Proceedings of the Ninth USENIX Security Symposium. pp. 85–96, URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.9679.

Rastogi, V., Suciu, D., Hong, S., 2007. The boundary between privacy and utility in data publishing. In: 33rd International Conference on Very Large Data Bases, VLDB 2007 - Conference Proceedings. pp. 531–542, URL: http://arxiv.org/abs/cs/0612103. arXiv:cs/0612103.

Reimers, N., Gurevych, I., 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, Online, pp. 4512–4525, doi:10/gm3nw2. URL:10/gm3nw2.

Rodina, Y., Westergaard, M., 2015. Grammatical gender in Norwegian: Language acquisition and language change. J. Ger. Linguist. 27 (2), 145–187, doi:10/gnh698. URL: https://www.cambridge.org/core/journals/journal-of-germanic-linguistics/article/abs/grammatical-gender-in-norwegian-language-acquisition-and-language-change/7185CA1B78B2F2966C7BD091B6D213D1#. Publisher: Cambridge University Press.

Rönnqvist, S., Kanerva, J., Salakoski, T., Ginter, F., 2019. Is multilingual BERT fluent in language generation? In: Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing. Linköping University Electronic Press, Turku, Finland, pp. 29–36, URL: https://aclanthology.org/W19-6204.

Rose, T., Haddock, N., Tucker, R., 1997. The effects of corpus size and homogeneity on language model quality. In: Proceedings of the Fifth Workshop on Very Large Corpora. Goldsmiths, University of London, London, pp. 178–191, URL: https://research.gold.ac.uk/id/eprint/30109/.

Rosenfeld, E., Ravikumar, P.K., Risteski, A., 2021. The risks of invariant risk minimization. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=BbNIbVPJ-42.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., Gurevych, I., 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp. 3118–3135, doi:10/gm6jhg. URL:10/gm6jhg.

Sasano, R., Kawahara, D., Kurohashi, S., 2009. The effect of corpus size on case frame acquisition for discourse analysis. In: Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Boulder, Colorado, pp. 521–529, URL: https://aclanthology.org/N09-1059.

Shadbolt, N., O'Hara, K., De Roure, D., Hall, W., 2019. Privacy, trust and ethical issues. In: Shadbolt, N., O'Hara, K., De Roure, D., Hall, W. (Eds.), The Theory and Practice of Social Machines. In: Lecture Notes in Social Networks, Springer International Publishing, Cham, pp. 149–200. http://dx.doi.org/10.1007/978-3-030-10889-2_4.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S., 2018. Generalizing across domains via cross-gradient training. arXiv:1804.10745 [cs, stat]. URL: http://arxiv.org/abs/1804.10745.

Shokri, R., Shmatikov, V., 2015. Privacy-preserving deep learning. In: Proceedings of the ACM Conference on Computer and Communications Security. Vol. 2015-Octob, pp. 1310–1321. http://dx.doi.org/10.1145/2810103.2813687, ISSN: 15437221.

Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. Membership inference attacks against machine learning models. In: Proceedings - IEEE Symposium on Security and Privacy. Institute of Electrical and Electronics Engineers Inc., pp. 3–18. http://dx.doi.org/10.1109/SP.2017.41, arXiv:1610.05820 ISSN: 10816011.

Søgaard, A., Ebert, S., Bastings, J., Filippova, K., 2021. We need to talk about random splits. arXiv:2005.00636 [cs]. URL: http://arxiv.org/abs/2005.00636.

Song, C., Raghunathan, A., 2020. Information leakage in embedding models. In: Proceedings of the ACM Conference on Computer and Communications Security. pp. 377–390. http://dx.doi.org/10.1145/3372297.3417270, URL: http://arxiv.org/abs/2004.00053. arXiv:2004.00053 ISSN: 15437221.

Song, C., Ristenpart, T., Shmatikov, V., 2017. Machine learning models that remember too much. In: Proceedings of the ACM Conference on Computer and Communications Security. pp. 587–601. http://dx.doi.org/10.1145/3133956.3134077, URL: http://arxiv.org/abs/1709.07886. arXiv:1709.07886 ISSN: 15437221.

Sousa, S., Kern, R., 2022. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. Artif. Intell. Rev. 1–66.

Stahlberg, D., Braun, F., Irmen, L., Sczesny, S., 2007. Representation of the sexes in language. In: Fiedler, K. (Ed.), Social Communication. Psychology Press, pp. 163–187, Num Pages: 25.

Sun, X., Wang, H., Zhang, Y., 2012. On the identity anonymization of high-dimensional rating data. In: Concurrency Computation Practice and Experience. Vol. 24, John Wiley & Sons, Ltd, pp. 1108–1122. http://dx.doi.org/10.1002/cpe.1724, URL: http://doi.wiley.com/10.1002/cpe.1724. Issue: 10 ISSN: 15320626.

Taulé, M., Martí, M.A., Pardo, F.M.R., Rosso, P., Bosco, C., Patti, V., 2017. Overview of the task on stance and gender detection in tweets on catalan independence. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages. IberEval 2017, pp. 157–177, URL: https://www.semanticscholar.org/paper/Overview-of-the-Task-on-Stance-and-Gender-Detection-Taul%C3%A9-Mart%C3%AD/502d59dcc6a7963190d7c1ea29d7bfc4f17e57e1.

Touileb, S., Øvrelid, L., Velldal, E., 2020. Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. In: Costa-jussà, M.R., Hardmeier, C., Radford, W., Webster, K. (Eds.), Proceedings of the Second Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain (Online), pp. 125–138, URL: https://aclanthology.org/2020.gebnlp-1.11.

Turc, I., Chang, M.-W., Lee, K., Toutanova, K., 2019. Well-read students learn better: On the importance of pre-training compact models. http://dx.doi.org/10.48550/arXiv.1908.08962, URL: http://arxiv.org/abs/1908.08962. arXiv:1908.08962 [cs].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Long Beach, California, USA, pp. 6000–6010, URL: https://dl.acm.org/doi/10.5555/3295222.3295349.

Verhoeven, B., Daelemans, W., 2014. CLIPS stylometry investigation (CSI) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3081–3085.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S., 2019. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076 [cs]. URL: http://arxiv.org/abs/1912.07076.

Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neurocomputing 312, 135–153, doi:10/gf6m39. URL: https://www.sciencedirect.com/science/article/pii/S0925231218306684.

Wang, X., Li, L., Ye, W., Long, M., Wang, J., 2019. Transferable attention for domain adaptation. Proc. AAAI Conf. Artif. Intell. 33 (01), 5345–5352, doi:10/ghkjzj. URL:10/ghkjzj. Number: 01.

Wang, Y., Wu, X., Hu, D., 2016. Using randomized response for differential privacy preserving data collection. In: Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference. Vol. 1558, Bordeaux, France, ISSN: 16130073.

Woller, L., Hangya, V., Fraser, A., 2021. Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 41–50, URL: https://aclanthology.org/2021.mrl-1.4.

Xu, Z., Aggarwal, A., Feyisetan, O., Teissier, N., 2020. A differentially private text perturbation method using regularized mahalanobis metric. In: Proceedings of the Second Workshop on Privacy in NLP. pp. 7–17. http://dx.doi.org/10.18653/v1/2020.privatenlp-1.2, URL: http://arxiv.org/abs/2010.11947. arXiv:2010.11947.

Xu, Q., Qu, L., Xu, C., Cui, R., 2019. Privacy-aware text rewriting. In: INLG 2019 - 12th International Conference on Natural Language Generation, Proceedings of the Conference. Association for Computational Linguistics (ACL), pp. 247–257. http://dx.doi.org/10.18653/v1/w19-8633, URL: https://www.aclweb.org/anthology/W19-8633.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., Kurzweil, R., 2020. Multilingual universal sentence encoder for semantic retrieval. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Online, pp. 87–94, doi:10/gmddgj. URL:10/gmddgj.

Zare, S., Nguyen, H.V., 2022. Removal of confounders via invariant risk minimization for medical diagnosis. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. In: Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, pp. 578–587. http://dx.doi.org/10.1007/978-3-031-16452-1_55.

Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, Association for Computing Machinery, New York, NY, USA, pp. 335–340, doi:10/gftmfb. URL: https://doi.org/10.1145/3278721.3278779.

Zhang, J., Li, W., Ogunbona, P., Xu, D., 2020. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. ACM Comput. Surv. 52 (1), 1–38, URL: https://dlnext.acm.org/doi/abs/10.1145/3291124.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.-W., 2019. Gender bias in contextualized word embeddings. 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, pp. 629–634, doi:10/ghcv6n. URL:10/ghcv6n.

Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.-W., 2018. Learning gender-neutral word embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp. 4847–4853, doi:10/ggwczw. URL:10/ggwczw.