Enhancing text anonymization via re-identification risk-based explainability

Benet Manzanares-Salor[a], David Sánchez[a]

[a]*Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Av. Paisos Catalans 26, Tarragona, 43007, Spain*

Abstract

Text anonymization is a challenging task usually carried out by human annotators, thereby incurring significant economic and temporal costs. Even though automated approaches have been proposed to mitigate those costs, practical mechanisms for text anonymization mostly rely on named entity recognition (NER), which is acknowledged to offer insufficient privacy protection. To tackle this issue, we propose a methodology to enhance the privacy protection attained by any text anonymization mechanism –with a focus on NER-based ones–, while providing empirical guarantees against re-identification rooted on the k-anonymity privacy model. Our method relies on a neural language model trained on the aggregated background knowledge that can be leveraged to conduct re-identification attacks. Then, it employs explainability techniques to detect and iteratively mask the unprotected terms that caused the greatest re-identification risk until a user-defined k-anonymity level is reached. On the contrary to most existing methods in the text anonymization literature, our approach allows to intuitively configure the desired level of protection, and to tune the trade-off between privacy and data utility preservation. Experiments show that our method is able to significantly and consistently lower the re-identification risk of NER-based anonymizations, and to compete against more sophisticated state-of-the-art text anonymization methods while being free of their costs and external dependencies.

Keywords: Privacy protection, Neural language models, Explainability,

Email addresses: benet.manzanares@urv.cat (Benet Manzanares-Salor), david.sanchez@urv.cat (David Sánchez)

1. Introduction

In the contemporary landscape of data-driven societies, scientific and business tasks require vast quantities of information for fueling innovation and research [23]. One of the most common and information-dense data sources is text. However, text data often contains sensitive information about individuals, as in the case of medical reports [48] or court cases [7]. Sharing or releasing such textual datasets therefore poses a direct conflict with privacy regulations, including, but not limited to, the European General Data Protection Regulation (GDPR) [14]. For instance, in the aforementioned medical reports, their direct release could expose highly confidential patient information (e.g., sensitive diagnoses), which would adversely affect individuals' standing in society (e.g., personal relationships or employment opportunities).

To circumvent privacy issues, data holders can resort to anonymization. That is, the irreversible masking of the data so that they can no longer be attributed to the subjects they refer to, thus not negatively impacting their lives.

Anonymization of structured data (i.e., tabular datasets) is a mature field, with multiple privacy models (such as k-anonymity [39] or ϵ-differential privacy [12]) that offer formal guarantees. Nevertheless, the anonymization of unstructured data such as text is a far more intricate task. Textual data, by its nature, is rich in ambiguity and context-dependence. Therefore, extensive domain knowledge is needed just to detect what information may facilitate re-identification [24, 8]. Due to these difficulties, text anonymization in the real world is usually carried out manually [4].

Some mechanisms have been proposed to mitigate the arduous work of manual anonymization. They mostly consists in preliminary annotations of potentially sensitive information –which are grounded on Named Entity Recognition (NER) tools– with the aim of decreasing the volume of manual annotations [37, 35]. These methods use machine learning models to detect predefined named entity types that are assumed to facilitate re-identification (e.g., names of people, locations or organizations).

However, NER-based text anonymization presents multiple drawbacks. First, as many authors have argued [40, 19, 24, 8, 37, 35], masking only predefined types of named entities does not sufficiently reduce the re-identification

risk. This is because identifying information in textual resources is not limited in type, so any term can be disclosive. This deficiency has been observed in various empirical evaluations of risk [28, 36, 34], which are needed because NER-based methods fail to provide any kind of privacy guarantee. Moreover, privacy requirements (that is, what to protect and up to which level) cannot be determined by the data holders because NER-based approaches systematically mask all detected entities regardless of the subject to whom they refer. Due to all these reasons, NER-based methods are usually considered as a starting point for human annotators to build upon [37, 35], rather than full-fledged anonymization solutions.

1.1. Contributions and plan

Considering the limitations of NER for text anonymization, in this paper we propose PETRE (*Privacy Enhancement for Text via Risk-oriented Explainability*), a privacy-grounded method that is able to improve NER-based anonymizations by enforcing a user-defined reduction of the empirical re-identification risk.

Our method brings the following contributions:

- Its risk assessment raises awareness to its users on the (usually non-negligible) residual disclosure risk of well-established and widely used text anonymization mechanisms, specifically, NER-based ones.
- Even though it has been conceived for NER-based approaches, it is general enough to be applied to any starting method, regardless of its operating principle.
- On the contrary to NER-based methods, our approach empowers the user to intuitively configure the desired level of privacy by specifying the subjects to be protected and the maximum re-identification risk allowed for those subjects. Our method is grounded on probabilistic k-anonymity [45] –a relaxation of the indistinguishably guarantee of the well-known k-anonymity privacy model [39]–, which restricts the re-identification probability to be lower or equal than $1/k$.
- In contrast with the standard application of k-anonymity to structured datasets, in our method the aggregated background knowledge considered to assess disclosure risks can be made subject-specific. Therefore, anonymization can be fine-tuned to each individual to be protected, which reduces unnecessary masking.

- Anonymization is enforced in an iterative manner, by which the empirical re-identification risk is re-evaluated after each masking until reaching the user-specified privacy threshold. This also contributes to minimize unnecessary masking.

- Because the anonymization of each document is guided by the observed risk (rather than employing a predefined masking criterion such as NER-based methods), the anonymized outcomes will present an empirical re-identification risk lower or equal than $1/k$ regardless of their content and the starting anonymization method. Thus, no *a posteriori* manual evaluations of the attained level of privacy would be required.

- In addition to defining the privacy requirements for anonymization, our method does not require human intervention and operates fully automatic and self-controlled.

To demonstrate the effectiveness of our proposal, we conducted several experiments on a general-purpose dataset of Wikipedia biographies [35] with a variety of starting anonymization methods encompassing not only NER-based approaches, but also more sophisticated privacy-grounded methods. We also show the generalizability of our approach in a domain-specific dataset of legal documents [37]. Results show that our approach consistently and proportionally reduces the re-identification risk of all the starting anonymizations, and competes against state-of-the-art methods while being free from their complexities, costs and external dependencies.

The rest of the paper is organized as follows. Section 2 provides background on data anonymization. Section 3 discusses related works on text anonymization. Section 4 presents and formalizes our approach. Section 5 details the implementation of our method. Experiments and results are reported in Section 6. Finally, Section 7 provides the conclusions and depicts several lines of future research.

2. Background

Anonymization consists in irreversibly masking a dataset to preclude the identification of the referred entities, such as individuals or organizations. Because adequately anonymized data cannot be unequivocally attributed to specific individuals, it is no longer considered personal and, therefore, outside the scope of privacy regulations such as the GDPR [14]. The data

distortion (or removal) introduced by masking necessarily impairs the informativeness/utility of the protected outcomes proportionally to the level of privacy attained. This defines a trade-off between privacy protection and data utility preservation that anonymization methods strive to optimize.

In the context of anonymization, data values can be classified into the following types:

- *Identifiers*: They allow to unequivocally identify the entity they refer to. Examples of identifiers include the complete name of an individual or IDs.

- *Quasi-identifiers*: Even though they do not enable unequivocal re-identification in isolation, they can when evaluated in combination with other quasi-identifiers, because those combinations are unequivocally associated to specific individuals in public identified sources (e.g., electoral rolls). The public sources that can be leveraged for re-identification are known as *background knowledge*. Examples of quasi-identifiers are zip code, age or job name.

- *Confidential values*: Sensitive data about the entity to be protected that should not be attributed to it. Under the scope of the GDPR[14] these include religious or political opinions, sexual orientation or healthcare information.

- *Non-confidential data*: they neither contribute towards re-identification nor disclose sensitive information. They can be either general terms (e.g., stop words) or information not related to the entity of interest.

On these basis, utility-preserving anonymization consists in removing all identifiers, masking the minimum amount of quasi-identifiers needed to prevent re-identification, and preserving confidential and non-confidential values. Truthful masking (that does not alter the meaning of the protected outcomes) can be done via suppression or generalization of values.

Anonymization should be ideally enforced under the umbrella of a privacy model, which provides clear privacy guarantees on the protected outcomes. In particular, the well-known k-anonymity model requires each individual in the anonymized dataset to be indistinguishable from at least $k-1$ other individuals [39] with respect to their quasi-identifying values. As a result k-anonymity upperbounds the re-identification probability to that of Eq.1.

By setting the value of k, users can tune the trade-off between privacy and utility preservation.

$$Re - identification\ probability \leq \frac{1}{k} \tag{1}$$

For structured data, such as relational databases, k-anonymity relies on an *a priori* classification of attributes (i.e., database columns) into the aforementioned types. Then, identifying columns are suppressed and quasi-identifying ones are suppressed or generalized to a common value in groups of size k or larger. This makes individuals indistinguishable in groups of k with respect to those quasi-identifying values.

The same guarantee against re-identification offered by k-anonymity can also be attained through probabilistic k-anonymity [45], a relaxation that removes the indistinguishability requirement but keeps limiting the re-identification risk to that shown in Eq.1. In practice, this means that records do not need to be indistinguishable with respect to their quasi-identifiers as long as the desired probability of re-identification is met. This opens the door to use masking operations other than suppression or common-ground generalizations.

ϵ-Differential privacy [12] is another widespread privacy model that offers a robust privacy guarantee. However, because it was designed for protecting interactive queries to a remote database, it is not well-suited for unconstrained data releases [10], which are the standard scenario tackled in text anonymization [24]. Moreover, because it relies on (perturbative) noise addition for masking, it hampers the truthfulness of the data[1].

3. Related work

As outlined in Section 1, text anonymization is a far more intricate task than the anonymization of structured data. The nature of language, characterized by its inherent ambiguity, nuances, complex semantics and reliance on context, renders the accurate detection of the (quasi-)identifiers to be masked an unresolved challenge [24, 8].

The most well-established approaches for (semi-)automatic text anonymization are those based on NER [22, 30, 50, 18, 33, 51, 20, 26, 5, 36]. They sim-

[1]Keeping anonymized outcomes truthful at record/individual level is a common requirement in text data releases, e.g., when releasing clinical documents [37]

plify the problem by assuming that (quasi-)identifiers correspond to certain types of named entities (e.g., dates or names of people). Then, NER-based anonymization consists in leveraging a machine learning model trained to detect those named entity types and replacing the detected ones by their types (e.g., PERSON, LOCATION or DATE). In very recent works, researchers started leveraging the exceptional NER capabilities of large language models (LLMs) [5, 44] for anonymization purposes.

Although applying NER-based anonymization to any corpus (or even a single document) is straightforward, it also presents multiple limitations [40, 19, 24, 8, 37, 35]. First, the assumption that (quasi-)identifiers exclusively correspond to specific types of named entities is fundamentally flawed. It is well-known in the privacy literature that *any* term or combination of terms –without being restricted by type– could potentially serve as a (quasi-)identifier, as long as it is related to the individual to be protected and appears in identified public sources [21]. Consequently, NER-based anonymization usually leaves many (quasi-)identifying terms unmasked (i.e., those not being named entities), thereby resulting in significant disclosure risks [24, 28, 36, 34]. Secondly, NER-based protection masks the pre-defined types of named entities independently of the data holder's privacy requirements and the subject to be protected. On the contrary to all privacy models and methods in the literature, on privacy in structured databases [21], NER-based methods do not define parameters to set the desired level of privacy (or to configure the trade-off between privacy and utility).

Alternatively to NER-based approaches, there exist some text anonymization methods that are better aligned with the principles of the Privacy-Preserving Data Publishing (PPDP) field [46, 15, 40, 41, 32, 19, 35]. On the contrary to NER techniques, these approaches focus masking on the subjects to be protected and are not restricted to predefined entity types. As a result, they offer more robust protection [24, 28].

Most of these methods rely on analyzing the correlation or mutual information between the terms appearing in the documents and the subjects to be protected. The smaller the set of correlated individuals, the more disclosive/identifying the terms will be. Approaches differ in how these correlations are obtained. Several authors estimate them from the co-occurrence of terms and identities in the Web [46, 40, 41]. Others gather statistics from structured databases, such as Wikidata [35]. Finally, in [19] the authors train a word embedding model on the documents to be protected, and use the similarity between the obtained embeddings as a measure of correlation. Nonetheless,

all these methods face complexity and scalability issues, caused by either the cost of performing Web searches, creating an in-domain structured database or training a language model from scratch. Therefore, they are much more difficult to apply in practice than NER-based approaches.

There are also some less related methods grounded on reinforcement learning [32] or differential privacy [15] that transform the documents into text embeddings or bags-of-words, and perturb them by using a neural model or noise. Nevertheless, these approaches are not well-suited for unconstrained data releases, because they do not preserve the textual format and readability of the source documents, and the random perturbation they add hampers the truthfulness of the information.

4. Privacy Enhancement for Text via Risk-oriented Explainability

As discussed in the previous section, even though NER-based methods are privacy lacking, they are still the only truly practical approaches to text anonymization. To tackle their limitations, in this section we present PETRE (*Privacy Enhancement for Text via Risk-oriented Explainability*), a PPDP-oriented general framework that takes the (privacy-lacking) anonymizations produced by NER and automatically and iteratively enhances them until reaching a user-defined privacy level, which is intuitively defined in terms of re-identification risk.

More specifically, our approach is able to enforce probabilistic k-anonymity [45] on top of the starting NER-based anonymizations, thereby attaining an intuitive and user-configurable privacy guarantee. For this, the user will specify the desired value of k and the background knowledge for the individuals to be protected. Then, our method will iteratively enhance the starting anonymizations until the observed probability of re-identification of each document to be protected is lower or equal than $1/k$.

This probability is assessed through a re-identification model that leverages the stated background knowledge to infer the individuals referred in the documents to be anonymized. Our approach then uses explanations from this re-identification model -specifically, how much each term in the document contributes to re-identification– to iteratively mask the most disclosive terms. This incremental process is repeated until each document meets the probabilistic k-anonymity requirement. By targeting only the terms that increase the re-identification risk, and by re-evaluating the residual risk at each masking iteration, our method avoids over-masking by design.

The architecture and workflow of our proposal are detailed in Section 4.1, while Section 4.2 formalizes it.

4.1. Architecture

In the following we provide a high-level overview of the PETRE framework by introducing its architecture, core components and workflow. These elements are depicted in Figure 1, with its core components highlighted in blue and bold text:

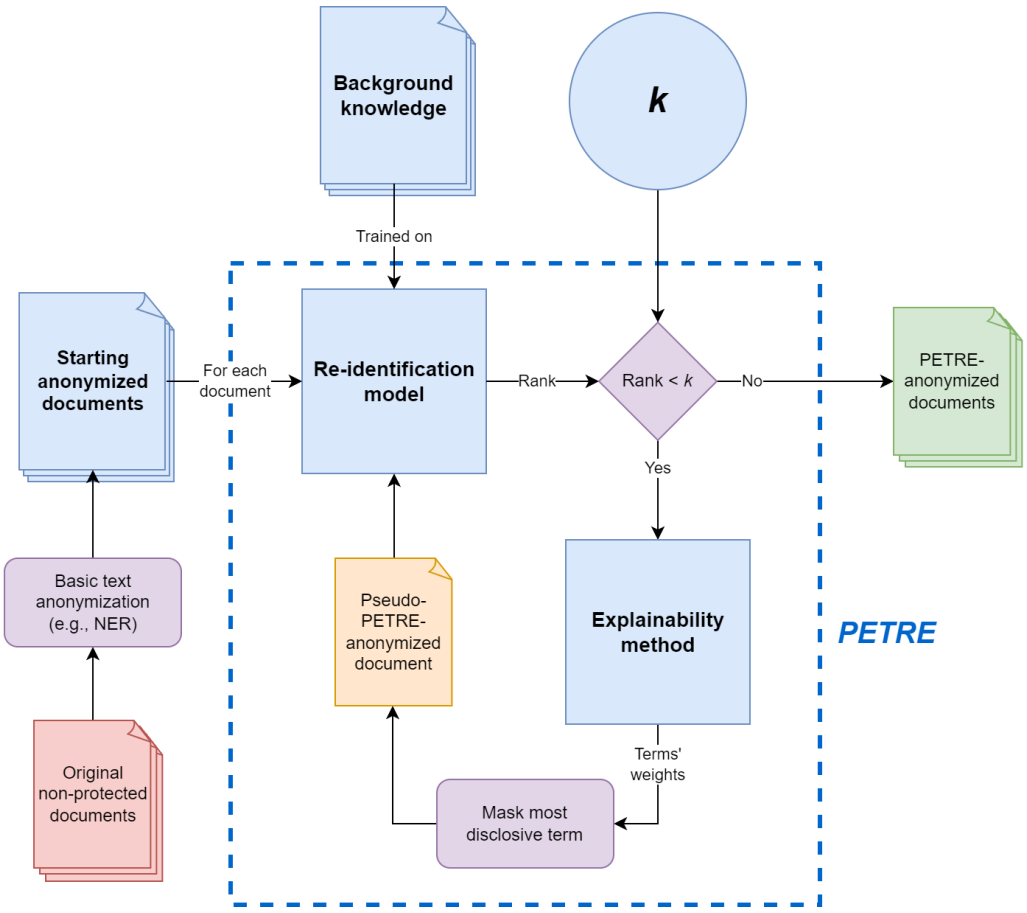


Figure 1: PETRE architecture and workflow

- **Starting anonymized documents**: The set of documents to be privacy-enhanced by PETRE. They are the result of the application

of a basic anonymization mechanism on the original (non-protected) documents. Each document (e.g., a medical record or a CV) is labeled with the name of the individual whose identity is to be concealed. Our approach can be applied to any method that suppresses or generalizes text spans, but it is specifically designed to improve NER-based anonymizations. This is because it addresses their main weakness, that is, poor privacy due to their limited coverage of entity types. On the other hand, PETRE's performance will be constrained by starting anonymizations that might provide higher privacy protection at the cost of over-masking. This is because PETRE cannot recover unnecessary masking incurred by the starting anonymization method.

- **Background knowledge**: A collection of (assumed) publicly available identified documents containing information on the individuals to be protected. These documents are used for training a re-identification model (defined below), which is used to assess the practical re-identification risk of the documents to be protected. The higher the overlap between the individuals' information in this background knowledge and that present in starting anonymized documents, the higher the re-identification risk will be because more (quasi-)identifiers can be leveraged to conduct re-identification attacks. In turn, a higher re-identification risk will make anonymization stricter, as more (quasi-)identifiers will require to be masked. A straightforward way to define this background knowledge is to consider the original non-protected documents as background knowledge. This constitutes a worst-case scenario, which will result in the strictest privacy protection (at the cost of utility). For a better balanced privacy/utility trade-off, data holders may carefully define which background knowledge could be realistically available for each individual. In particular, the scarcer the background knowledge for an individual, the less masking (and subsequent information/utility loss) is required for achieving the desired level of privacy protection.

- **Re-identification model**: A classifier trained on the aforementioned background knowledge to predict the individuals' identity. This component evaluates the privacy level of a (protected) document. More specifically, for any given document, it predicts a probability distribution over the set of possible individuals to be re-identified. PETRE

uses the produced rank –that is, the position of the corresponding individual's identity within the predicted probability distribution– to assess the re-identification risk: the lower the rank, the higher the probability of successful re-identification and, therefore, the higher the risk.

- k: A positive integer that defines the desired level of privacy in terms of maximum allowed risk of re-identification ($1/k$, as stated by probabilistic k-anonymity). PETRE uses it as a reference for the rank produced by the re-identification model for each document to be protected. In this way, PETRE can ascertain whether a document necessitates further masking ($rank < k$) or if it has already reached the desired level of privacy ($rank \geq k$). In other words, it defines the ending condition of the iterative anonymization. The desired level of protection, as well as the balance between privacy and utility, is determined by this parameter. Users should select k according to the characteristics of the dataset being anonymized, and the applicable regulatory requirements. First, the value of k should be proportional to the dataset's size. For instance, smaller datasets with only a few hundred samples, would not allow choosing a very large k without a significant loss of data utility. Second, the acceptable level of re-identification risk should correspond to the intended purpose of the data. While any $k \geq 2$ guarantees that no individual can be uniquely re-identified, as required by Recital 26 of the GDPR[14][2], lower re-identification risks are usually desired. For example, as highlighted in [13], a re-identification risk of approximately 0.09% or less (equivalent to $k > 10$) is considered reasonable, as it has been explicitly proposed by two independent health regulators [2, 6].

- **Explainability method**: A module capable of estimating the influence of the document's terms in the probability predicted by the re-identification model for the individual it refers to. The output of this component are the terms' weights in the re-identification success. Weights' positive or negative magnitudes indicate their influence (favorable or unfavorable, respectively) on the classification probability of the individual. PETRE uses these weights to mask the terms with the greatest influence in the re-identification risk at each iteration. In terms of explainability, this module obtains a local explanation for the

[2] `https://gdpr-info.eu/recitals/no-26/`

probability predicted for a specific class/individual. Note that, for our application, the fidelity of the weights takes precedence over their interpretability, as our approach is automatic and does not involve human interpretation. Valid explainability methods can be model-agnostic or those model-specific that are compatible with the design of the re-identification model (i.e., a classifier). Model-agnostic approaches grounded on additive feature attribution such as SHAP [25] or LIME [38] are adequate.

PETRE's design is general and allows each of its functional components (i.e., the re-identification model and the explainability method) to be replaced, e.g., to improve their performance as new developments arise. Together with the possibility of configuring the anonymization parameters (i.e., k and background knowledge) for specific privacy/utility needs, this defines a more flexible and customizable approach than existing text anonymization methods while still keeping it practical.

Now that the core components have been elucidated, we can outline the workflow depicted in Figure 1. Each document from the **starting anonymized set** is processed individually by PETRE. First, privacy level of the document is assessed by the **re-identification model**, which has been previously trained with the **background knowledge**. If the rank predicted by the re-identification model is lower than the specified $\boldsymbol{k}$, the document will require further masking. In this case, the **explainability method** estimates the weights for all the terms in the document. Subsequently, the term with the largest weight is masked (e.g., removed), thereby incrementally lowering the re-identification risk. Afterwards, the document is re-evaluated for re-identification risk and the process is repeated until the document reaches or surpasses the specified k. This workflow is fully automatic and self-controlled, thereby requiring no human intervention.

4.2. Formalization

Let $A = \{(D_1, I_1), \ldots, (D_n, I_n)\}$ be the set of starting anonymized documents, consisting on n pairs of protected documents D and the identities of the corresponding individuals to be protected I.

Let $\Gamma(d)$ be the re-identification model, which receives a document d and outputs a probability distribution $P = (p_1, \ldots, p_n)$ summing one. Each p_i represents the probability that d actually refers to the individual I_i. This machine learning model is trained with a background knowledge set $B =$

$\{(C_1, I_1), \ldots, (C_n, I_n)\}$, mapping the set of public identified documents C to their corresponding individuals in I.

For a given document-individual pair (D_i, I_i) from A, the document D_i is considered unequivocally re-identified if the probability p_i (corresponding to I_i) returned by $\Gamma(D_i)$ is the highest in the distribution P. This is in line with the well-known record linkage attack [1], which data anonymization aims to prevent. Subsequently, we can assess the re-identification risk of a document D_i according to the ranked position of p_i in the distribution P. We denote $Rank(d, i, \Gamma)$ as the function that, given a document d, an individual i and the re-identification model Γ, returns that ranked position. For example, $Rank(d, i, \Gamma) = 1$ indicates that $\Gamma(d)$ returned the largest probability for the individual i. On this basis, the upper-bound for re-identification risk can be intuitively derived as follows:

$$Rank(d, i, \Gamma) = k \implies Re - identification\ risk(d) \leq 1/k \tag{2}$$

This way of measuring risk is in line with the privacy guarantee of k-anonymity (see Eq.1). To be more precise, a $rank = k$ conforms to the principles of probabilistic k-anonymity [45]; that is, it limits the re-identification risk to $1/k$ without requiring the documents of the k individuals be indistinguishable from each other –as stated by standard k-anonymity–. It is important to note that, in text anonymization, indistinguishability is neither necessary nor advantageous, as it only hampers utility by requiring to resort in common generalizations across different documents.

Let $k \in \mathbb{N}^+$ be the privacy threshold, which defines the minimum rank acceptable for a PETRE-anonymized document. For instance, a value $k = 1$ means that no anonymization is necessary, while $k = 2$ requires that all documents have, at least, a $rank = 2$. If a document is assessed with a rank lower than the desired k, it will need further masking.

The explainability method $\xi(d, i, \Gamma)$ receives a document d, an individual I_i and the re-identification model Γ. The document d consists in a set of terms $(t_1, \ldots, t_m)$, with m varying for each d. The output of ξ are the weights($\in \mathbb{R}$) of each document term $(w_1, \ldots, w_m)$ for the probability p_i predicted for the individual I_i. Positive weights indicate a direct influence (the higher, the more the influence) of the corresponding terms towards re-identifying I_i; that is, terms with positive weights are effectively acting as (quasi-)identifiers for I_i. On the other hand, negative weights correspond to terms that hinder the re-identification, whereas weights close to zero are ex-

pected for non-meaningful terms (such as stop words) or terms non-related to the individual to be protected. We operate under the assumption that the explainability method's fidelity might be imperfect, thereby resulting in weights that may not precisely represent the impact of masking the corresponding terms. Due to this reason, weights are recalculated after each masking iteration and disclosure risk re-evaluation. In this way, we ensure that they end up being as faithful as possible, even if this means conducting more iterations.

Algorithm 1 formalizes our iterative privacy-enhancement method. First, the re-identification model Γ is trained by using the background knowledge B (line 2). Then, for each document of the starting anonymized set A (line 3), its rank is measured (line 4) and, as long as it exceeds the privacy threshold k (line 5), the document is iteratively masked (lines 6 to 9). For each masking step, first, the explainability method is leveraged to obtain the terms influences' weights in re-identification (line 6). Afterwards, in line 7, the index of the term that contributed the most to re-identification (i.e., that with the maximum weight) is obtained. That index is then employed in line 8 to mask the corresponding term. Masking only the most disclosive term at each step aims to minimize the amount of masked terms, which will contribute to retain data utility. At the end of the masking loop (line 9), the rank of the document is re-evaluated. Once the document is sufficiently anonymized, it is stored in A^{PETRE} (line 11), which is the PETRE-enhanced set of documents produced by the algorithm (line 13).

5. PETRE implementation

In this section, we provide a comprehensive account of the practical implementation of the general framework introduced above. Based on the architecture described in Section 4.1, we focus on the re-identification model and the explainability method. In the following, we determine and discuss the technologies used for both components. Algorithm 2 depicts the specific adaptations made to the general PETRE framework outlined in Algorithm 1 to accommodate the chosen solutions.

For the re-identification model, we considered the TRIA (Text Re-Identification Attack) method proposed in [28], which trains a language model-based classifier specifically for text re-identification. The model is trained on background knowledge through a two-phase process: additional pre-training (e.g., masked language modeling), and fine-tuning (e.g., supervised classification training).

Algorithm 1 PETRE

Require: A, B, Γ, ξ, k
1: $A^{PETRE} \leftarrow \emptyset$
2: $\Gamma \leftarrow Train(\Gamma, B)$
3: **for** $(document,\ individual) \in A$ **do**
4: $\quad rank \leftarrow Rank(document, individual, \Gamma)$
5: $\quad$ **while** $rank < k$ **do** ▷ Mask until privacy threshold is met
6: $\quad\quad weights \leftarrow \xi(document, individual, \Gamma)$
7: $\quad\quad termIdx \leftarrow ArgMax(weights)$
8: $\quad\quad document \leftarrow Mask(document, termIdx)$
9: $\quad\quad rank \leftarrow Rank(document, individual, \Gamma)$
10: $\quad$ **end while**
11: $\quad A^{PETRE} \leftarrow A^{PETRE} \cup document$
12: **end for**
13: **return** A^{PETRE}

During fine-tuning, the classifier is trained to identify the individuals corresponding to the raw documents in the background knowledge, and the anonymized versions of these documents. The idea is to teach the model to learn comprehensive information about the individuals while adapting to the transformations introduced by anonymization. Once trained, the model can be used to predict a probability distribution over individuals from the background knowledge when presented with a previously unseen anonymized document. TRIA perfectly matches the Γ function defined in Section 4.2, as it can be trained on background knowledge set B and infer a probability distribution P over the set of possible individuals I for an input document d. Consequently, in Algorithm 2 the parameter Γ has been renamed as $TRIA$.

It is worth noting that TRIA's predictions are done per document splits, and those are later aggregated to obtain the document-level probability distribution. This is done due to both the limited input length supported by the language model in which TRIA relies, and to improve the prediction accuracy. The latter is because aggregating partial predictions, as ensemble learning methods do, is usually more reliable than resorting on a single global prediction. For our implementation, we configured TRIA code[3] to define one

[3] `https://github.com/BenetManzanaresSalor/TextRe-Identification`

split per sentence, using default values for the other settings. The model then produces sentence-level logits, which are then summed and passed through a softmax function to compute the final probability distribution. As backbone pre-trained language model upon TRIA builds on, we leverage DistilBERT [43], a smaller version of the well-known BERT model that offers similar accuracy, but any other general model could be employed.

On the other hand, we adopted SHAP [25][4] as the explainability method. SHAP is a state-of-the-art model-agnostic approach that estimates the influence of each input feature in the prediction (local explanation) of a regressive model. In our framework, the re-identification model is a classifier, which in turn can be abstracted as a multi-regression method (i.e., one regression per class). Subsequently, SHAP can be used to measure the weight (i.e., influence) of each document's term (i.e., input feature) in the re-identification probability (i.e., regression value) of a specific individual (i.e., class). This aligns with the individual-specific explainability function defined as ξ in Section 4.2 and employed in Algorithm 1. As a consequence, Algorithm 2 uses $SHAP$ in place of ξ. However, making SHAP usable for PETRE requires from some modifications.

Firstly, a SHAP explanation entails performing multiple model predictions by randomly replacing combinations of input features (i.e., terms). The amount of combinations scales quadratically with the number of terms. This causes a scalability issue that, in conjunction with the potentially non-negligible cost of the re-identification model prediction, can lead to impractical execution times, especially when dealing with long documents. To address this problem, we leverage TRIA's particularity of processing documents per splits (prior to aggregation) and knowledge on SHAP foundations.

SHAP's need of evaluating features combinations comes from the assumption that the features influence may not be independent. That is, the influence of a feature may vary depending on other features of the same input. This is in line with the *compositionality* property of language, being each term's meaning (and subsequent influence in prediction) affected by its context. Nevertheless, TRIA processes documents by splits, so the terms from different splits cannot influence each other. This nullifies the necessity of evaluating their combinations. Therefore, SHAP can be applied to each split (rather than to the whole document prediction), thereby drastically reducing

[4] `https://github.com/shap/shap`

the number of term combinations to evaluate while not losing explanation quality. For instance, if having a document with m terms divided into s equally-lengthen splits, weights' computing cost is reduced from $O(m^2)$ to $O(s * (\frac{m}{s})^2)$.

Unfortunately, we observed that even this cost reduction may not be enough for avoiding unpractical runtimes. To tackle this issue, we recall that PETRE only uses the weights to ascertain the most disclosive terms of the document (see line 7 of Algorithm 1). Because it is expected that these terms are contained in the most disclosive split, the explanation can be limited to this split (i.e., lines 7 and 8 in Algorithm 2). This further reduces explanation cost from the aforementioned $O(s * (\frac{m}{s})^2)$ to $O((\frac{m}{s})^2)$.

Finally, it is relevant to establish a clear definition of what constitutes a term, particularly when it comes to the detection of (quasi-)identifiers. SHAP calculates its weights at the token-level[5], being tokens the input features for language models. However, in the context of text anonymization, the minimum meaningful units are typically words and/or noun chunks. This holds true for our framework as well, since attempting to mask individual tokens instead of complete words (e.g., "ina" in "Angelina") or noun chunks (e.g., "New" in "New York") would be ineffective. Hence, we designate both noun chunks and individual words (those not part of chunks) as terms. For noun chunks detection, we leverage the spaCy library[6]. Subsequently, we calculate terms' weights as the summation of the weights of the constituent tokens. In this way, masking becomes independent of the language model vocabulary, which is optimized for performance on NLP tasks rather than anonymization. Specifically, the vocabulary is based on tokens instead of terms, the latter being the relevant entities for anonymization.

6. Experiments

We next assess the performance of PETRE on multiple text anonymization mechanisms and datasets. All tests were conducted on the implementation described in Section 5 running on a computer equipped with a Nvidia 4080 GPU, an i7 10700 CPU, 32GB of RAM, and a 1TB NVME

[5] https://github.com/shap/shap/blob/master/notebooks/text_examples/translation/Machine%20Translation%20Explanations.ipynb

[6] https://spacy.io/usage/linguistic-features#noun-chunks

Algorithm 2 PETRE with TRIA and SHAP

Require: $A, B, TRIA, SHAP, k$
$A^{PETRE} \leftarrow \emptyset$
$TRIA \leftarrow Train(TRIA, B)$
for $(document,\ individual) \in A$ **do**
    $rank \leftarrow Rank(document, individual, TRIA)$
    while $rank < k$ **do** ▷ Mask until privacy requirement is met
        $splits \leftarrow Split(document)$
        $mdSplit \leftarrow GetMostDisclosiveSplit(splits, TRIA)$
        $weights \leftarrow SHAP(mdSplit, individual, TRIA)$
        $termIdx \leftarrow ArgMax(weights)$
        $mdSplit \leftarrow Mask(mdSplit, termIdx)$
        $rank \leftarrow Rank(document, individual, TRIA)$
    end while
    $A^{PETRE} \leftarrow A^{PETRE} \cup document$
end for
return A^{PETRE}

SSD. Code and data for all the reported experiments are available at `https://github.com/BenetManzanaresSalor/PETRE`.

6.1. Evaluation datasets

We leverage the general-purpose Wikipedia-based dataset presented in [35], which is available in a public GitHub repository[7]. It contains 553 abstracts of randomly selected Wikipedia biographies, with individuals ranging from football players to World War II generals. Wikipedia biographies are frequently employed as evaluation data for text anonymization due to the substantial and varied amount of (quasi-)identifying terms they contain, making them a challenging testbed [24, 40, 42, 46]. In this particular dataset, the abstracts are considered the sensitive text to be protected.

These documents have been manually annotated for anonymization by several human annotators –one NLP researcher and four undergraduate students in law– by following the guidelines described in [37]. Annotations encompassed the detection of disclosive terms, categorization into semantic groups (e.g., person, date, organization, etc.), and determining whether its

[7] `https://github.com/anthipapa/bootstrapping-anonymization`

masking was necessary for concealing individual's identity. For the first two tasks, annotations of the spaCy NER tool[8], but annotators were instructed to make any necessary corrections. Manual anonymizations were generated by replacing the terms marked for masking by their semantic category, as typical done in NER-based anonymization. The privacy goal provided to the annotators was in line with the standard for manual anonymizations, which involves masking the text to the extent that the corresponding individual becomes unrecognizable. The attainment of that goal is defined by the annotator's own judgment and common sense.

We also assess the ability of our approach to generalize by applying it to a domain-specific dataset. For this purpose, we used the Text Anonymization Benchmark (TAB) [37], which comprises 1,268 English documents derived from European Court of Human Rights (ECHR) case files. Concretely, the *Procedure* section of each court case, which is abundant in (quasi-)identifiers, was considered the sensitive text to be protected. Documents were manually annotated by 12 law students, who identified sensitive text spans and classified them into entity types, adhering to annotation guidelines very similar to those used for the Wikipedia-based dataset. For evaluation, we employed the test subset of this dataset, consisting of 127 documents with corresponding annotations.

6.2. Anonymization methods

We tested the performance of PETRE on a variety of state-of-the-art automatic anonymization methods. These include not only NER tools (which would be the natural target for PETRE), but also PPDP-oriented approaches. The latter will show the capability of PETRE to provide consistent results regardless of the type and operating principles of the starting anonymization. The approaches and tools described below have been used in the text anonymization literature as evaluation baselines [19, 28, 37]:

- **Stanford NER** [27]: A collection of three pre-trained NER models considering different sets of entity types: NER3, which detects PERSON, ORGANIZATION and LOCATION types; NER4, which is able to detect PERSON, ORGANIZATION, LOCATION and MISC (miscellaneous) types; and NER7, which detects PERSON, ORGANIZATION, LOCATION, MONEY, DATE, PERCENT and TIME.

[8] `https://spacy.io/api/entityrecognizer` were given as the starting point

- **Microsoft Presidio**[9]: A NER-based tool specifically designed for text protection. Presidio supports a wide diversity of entity types, encompassing both (quasi-)identifying and confidential data (e.g., credit card numbers). For the purpose of text anonymization, we selected those conveying quasi-identifying information: PERSON, LOCATION, NRP –persons nationality, religious or political group– and DATE_TIME.

- **Disclosure protection based on word embeddings** [19]: As introduced in Section 3, this method employs a word embedding model (Word2Vec [31]) trained from scratch on a corpus including both the documents to be protected and external documents providing general (if possible, in-domain) knowledge. The aim of the embedding model is to identify text spans that bear semantic similarity with the individuals requiring protection. Given that word embeddings encode the semantic connections between words, it is expected that the words that may facilitate re-identification would also be close to the individual's name in the embedding space. This is because both the name and those close terms tend to co-occur. Consequently, words with embeddings exhibiting a cosine similarity to the individual's name greater than a user-defined threshold are considered as (quasi-)identifiers and are subjected to masking. Although a more intricate ontology-based method is proposed for masking through fine-grained generalizations, for the sake of fair comparison (all other methods, including manual anonymizations rely on suppression or entity-type generalization), we evaluated this method as if sensitive terms were suppressed. We trained the word embedding model on the 553 biographies to be protected (including both abstracts and bodies) along with a corpus of 19k Wikipedia biographies, as done in the original paper [19]. We selected the same similarity threshold used by the authors in their original experiments (i.e., $t = 0.25$), which were also conducted on Wikipedia biographies. In the following, this anonymization will be referred to as *Word2Vec*. We employed the public GitHub implementation available at `https://github.com/fadiabdulf/automatic_text_anonymization`.

- **k-anonymity-based on knowledge graphs** [35]: An approach grounded on a knowledge graph encompassing the term distribution within

[9] `https://github.com/microsoft/presidio`

Wikidata[10]. Its purpose is to identify terms occurring in fewer than k documents. Following this, the method enforces k-anonymity [39] on the set of documents to be protected by iteratively suppressing terms until no combination of them is present in fewer than k documents. Term selection can be carried out using either a *greedy* approach, which prioritizes the masking of the least frequent term first, or in a *random* manner, where the next sensitive term to be masked is selected arbitrarily. The greedy strategy is anticipated to yield a lower (though not necessarily the minimum) number of masked terms, which should help retaining utility more effectively. In the following, these two versions of the method will be referred to as *k-anonymity greedy* and *k-anonymity random*. Because these methods were already tested on the same corpus of 553 abstracts we use here, we employed the results provided by the authors (with $k = 5$) "as is".

It is worth noting that both PPDP-oriented methods, although being more practice-oriented than most related approaches in the literature [37], would still be quite difficult to apply in a real-life scenario. On the one hand, Word2Vec requires training a word embedding model from scratch, which can incur in a substantial cost and requires the compilation of suitable external corpora to avoid overfitting. Moreover, because masking is based on an empirically-tuned similarity threshold, it does not offer a clear guarantee on how the documents have been protected. On the other hand, the method from [35] requires building a knowledge graph, which, for most domains, would imply bespoke construction by humans. The associated effort of creating such a knowledge graph on demand could rival or even exceed that of manual anonymization.

Given the iterative and self-adaptive design of PETRE, intuition suggests that the privacy improvement it enables will be inversely proportional to the quality of the starting anonymization. That is, the more privacy-lacking the starting anonymization is, the larger the enhancement will be. Under this assumption, we expect the protection of the (privacy-lacking) NER-based methods to improve substantially with PETRE. For PPDP-oriented approaches, which are far more complex in their application but have shown to provide much more robust empirical privacy [28], we expect more subtle improvements; but we still decided to include them in order to test the gen-

[10] `https://www.wikidata.org/`

erality of our approach.

Consistently with the masking strategies enforced by the methods above, the additional masking performed by PETRE consisted in *removing* sensitive terms from the documents.

6.3. Background knowledge

Because we use manual annotations as groundtruth, PETRE's background knowledge should closely align with what human annotators had access to. For both datasets, annotators likely considered the own documents to protect as their background knowledge. Therefore, it is reasonable to set the whole sets of 553 (Wikipedia) and 127 (TAB test) documents as background knowledge for PETRE. This setting also corresponds to a worst-case scenario for privacy, where we assume perfectly knowledgeable attackers. This background knowledge was utilized to train the re-identification model employed in the experiments. This training process, which consists in further pre-train and fine-tune the DistilBERT language model, and is done once for the whole set of documents to anonymize, took 104 minutes for the Wikipedia dataset and 121 minutes for the TAB test dataset.

6.4. Evaluation metrics

To gauge the improvement brought by PETRE over the baseline methods, we employed the standard evaluation methodology adopted in the text anonymization literature, which consists in comparing automatically anonymized outcomes to a manually anonymized groundtruth [22, 30, 50, 18, 33, 51, 20, 26, 40, 41, 19, 28, 37, 35]. The agreement between both anonymizations is measured through the IR-based metrics *recall* and *precision*.

Recall (Eq.3) measures the percentage of disclosive terms (according to the manual anonymization) that were also masked by the automatic method to be evaluated. Because missed terms are considered disclosive, this metric is understood as an indicator of the attained level of privacy protection.

$$Recall = \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives} \tag{3}$$

Precision (Eq.4) accounts for the percentage of terms masked by the automatic method that were actually needed to be masked according to the manual anonymization. Because incorrectly masked terms would unnecessarily hamper the readability and informativeness of the anonymized outcomes,

precision is understood as a measure of utility preservation.

$$Precision = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \tag{4}$$

F-score (Eq.5), is the weighted harmonic mean of precision and recall, and represents the overall anonymization quality considering both utility and privacy perspectives. The β factor determines the weight assigned to recall relative to precision, thereby reflecting the prioritization of privacy over utility.

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \tag{5}$$

Even though the F_1-score –which gives equal weight to both precision and recall– have been traditionally employed, several authors have recently advocated for the use of the F_2-score over the F_1-score [16, 29]. The reason is that recall should take precedence over precision in text anonymization because a low recall (regardless of the level of attained precision) could render anonymization ineffective.

All these annotations-based metrics have been calculated on a per-document basis, and the average across all the documents has been reported.

In addition the standard measures above, we also evaluated the extend to which the anonymized documents retained their analytical utility in a downstream task. We focused on document clustering, which relies on the semantic intra- and inter-similarity among documents. Analytically useful anonymized documents mean that clusters created from those documents closely align with those derived from the original texts, ensuring that similar insights can still be extracted. For clustering, we adopted the method described in [47], which uses BERT [9] (specifically the *bert-base-cased* model, available at `https://huggingface.co/google-bert/bert-base-cased`) to generate document embeddings. Embeddings were computed by averaging the token-level embeddings for each document, and clustered using the K-means++ algorithm [3] with $K = 4$. The choice of K was based on the optimal inertia achieved when clustering the original documents. To address the variability introduced by random initialization, clustering was repeated 50 times for each document set, and the result with the lowest inertia was selected. We compare the cluster similarity between original and anonymized documents using *Normalized Mutual Information* (NMI) score [17]. For robustness, the clustering process was repeated 5 times, and the average NMI

was reported. In total, each clustering configuration was executed 250 times (50×5) to minimize the influence of randomness.

6.5. Setting the privacy parameter k

The privacy protection attained by PETRE directly depends on the value of the k parameter, which states the desired level of (probabilistic) k-anonymity. To enable a fair comparison against the human anonymizations employed as groundtruth, in the following we try to ascertain which k value matches the best with the level of protection offered by those manual anonymizations.

For this, we measured the effective document rank provided by the re-identification model employed by PETRE for each manually anonymized document. As explained in Section 4.2, a $rank = k$ states an actual re-identification risk equal or lower than $1/k$ or, in other words, that the document is probabilistically k-anonymous. In particular, we measured the rate of documents that are compliant with 2-anonymity or above (that is, the percentage of documents that obtain a rank equal or higher than 2). As a reference, the percentage of original unprotected documents that were already 2-anonymous was just 1% for both datasets, which shows that the documents in those datasets are privacy challenging. Surprisingly, manual anonymizations got a 2-anonymous compliance rate around 19% for both datasets, which also means that 100%-19%=81% of the documents had a $rank = 1$ or, in other words, that a 81% of the manually anonymized documents can be unequivocally re-identified by the re-identification model.

The lack of privacy protection for manual anonymizations can be (at least partially) attributed to missed combinations of quasi-identifiers. While human annotators can proficiently identify and mask direct identifiers, such as names or addresses, they struggle recognizing the co-occurrence of multiple quasi-identifiers across documents. For example, if a document reveals that the individual to be protected is a male, plays football, comes from Spain, and has blond hair, an annotator will probably mask the term "Spain" based on the awareness that it is the sole individual from that country in the dataset. However, the annotator will more easily overlook the fact that this individual is also the only male footballer with blond hair in the dataset, especially if this information is scattered throughout the document. This oversight is significant, because it is precisely those unique combinations of quasi-identifiers that are exploited by re-identification attacks. Moreover, according to the annotation guidelines in [35] and [37], manual anonymization only aimed to preclude strict re-identification or, in other words, annotators

never aimed for a re-identification probability larger than $k = 2$. All in all, this illustrates the limitations of manual anonymization efforts –which are prone to errors and omissions [24]–, and defines a relatively privacy-lacking groundtruth, which will pose limitations to the interpretation of some of the evaluation metrics.

Given the above, we conclude that a $k = 2$ for PETRE provides the closest match to the level of protection achieved through manual anonymizations. Even though this value is still higher than the rank for most anonymized documents in the human-based groundtruth, it is the lowest one that offers a guarantee against re-identification.

6.6. Results

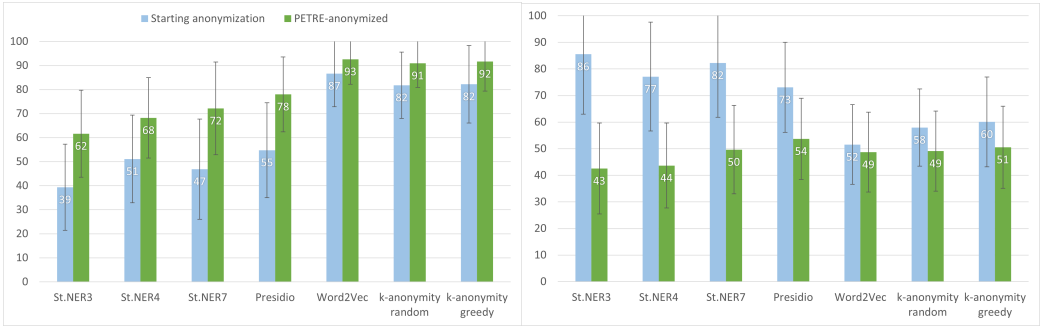
We next report a comprehensive evaluation of PETRE by using the Wikipedia-based dataset. First, we investigate the behavior of PETRE with respect to the different starting anonymizations. Afterwards, we analyze the influence of the k parameter and the background knowledge on PETRE's enhancements. In the last subsection, we evaluate the generalization capabilities of PETRE to the domain-specific TAB dataset.

6.6.1. By starting anonymizations

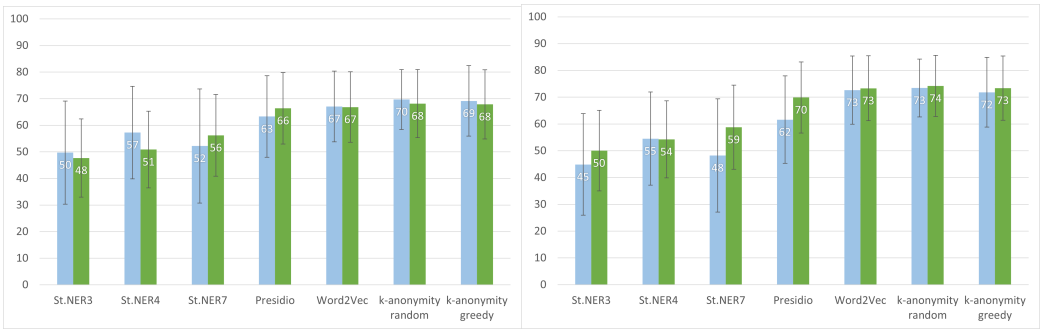
Figure 2 reports the evaluation metrics obtained by each starting anonymization introduced in Section 6.2 and its PETRE-anonymized counterpart. Error bars correspond to the standard deviation of the per-document metrics (precision, recall and F-scores).

PETRE proves highly effective in enhancing the recall (or, in other words, the privacy protection) of the starting anonymizations. As expected, the improvements are also larger for the most privacy-lacking methods, which are the NER-based approaches for which PETRE was designed. PPDP-based methods show minor improvements, but those are consistent across them. This can also be seen in the output examples depicted in Table 1, which correspond to the starting anonymizations with the lowest and highest recall: NER3 and k-anonymity greedy, respectively. Moreover, as reported in Figure 2, PETRE produced *consistently* improved anonymizations, with standard deviations being a 12% lower than for starting anonymizations on average for all annotations-based metrics.

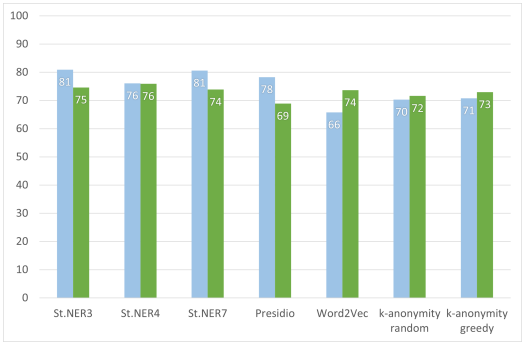
Recall improvements appear to be inversely proportional to the starting recall, as shown in Figure 3, which depicts the Pearson correlation between the starting recall and the recall improvement rate brought by PETRE.



(a) Recall (b) Precision



(c) F_1-score (d) F_2-score



(e) NMI

Figure 2: Evaluation metrics of starting anonymizations and their PETRE-anonymized counterparts

The negative correlation with an absolute magnitude > 0.9 ($p\text{-}value < 0.0003$) shows a very strong and statistically significant relationship between PETRE's improvement and the starting anonymization quality. This empir-

Table 1: PETRE-enhancing examples. Dashed underline corresponds to terms masked by the starting anonymization method, whereas solid underline shows additional terms masked by PETRE.

Method	Anonymization (underlined means masked)
St.NER3 PETRE-enhanced	Andrés Alejandro Palomeque González (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name Abismo Negro, which is Spanish for "Black Abyss", in the Asistencia Asesoría y Administración (AAA) promotion.
k-anonymity greedy PETRE-enhanced	Andrés Alejandro Palomeque González (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name Abismo Negro, which is Spanish for "Black Abyss", in the Asistencia Asesoría y Administración (AAA) promotion.

ical observation strongly reinforces the premise outlined in Section 4, suggesting that the worse the privacy of the starting anonymization, the more pronounced PETRE's enhancement will be. As a result, PETRE is capable of improving NER-based starting anonymizations to a level similar to that of PPDP-oriented approaches, all without the need for the costly and potentially inaccessible resources associated with the latter.

For precision, it is worth noting that it decreases to approximately 55% for all methods, regardless of the initial value. To properly interpret these results, we should recall the discussion in Section 6.5, in which we showed that most manually anonymized documents did not even fulfilled being 2-anonymous. Therefore, because PETRE was set to achieve 2-anonymity (a significantly stronger protection than that offered by the manual anonymization), it is expected that many of the terms masked by PETRE to achieve this level of privacy were not manually annotated, thereby counting as unnecessary maskings when computing precision. This fact is also shown by the nearly equivalent precision achieved by PETRE for all the starting anonymizations: in all cases, PETRE is masking the same set of (non-annotated) terms to reach the desired level of protection. As also discussed in Section 6.5,

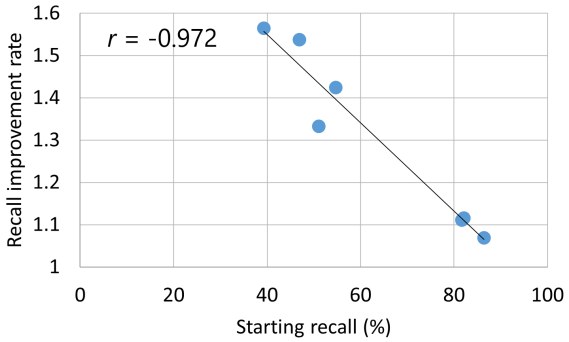


Figure 3: Correlation between the recall of starting anonymizations and the improvement rate brought by PETRE

these terms probably correspond to unique combinations of quasi-identifiers overlooked by human annotators. In any case, because it is not possible to distinguish between terms that *needed* to be masked to reach the desired privacy (i.e., *truly* true positives) from those that did not (i.e., *truly* false positives), precision results are largely inconclusive. It is worth noting that the precision of Word2Vec remains nearly unchanged after applying PETRE, whereas recall was improved by a non-negligible 6%. This indicates that most of the terms masked by PETRE (and overlooked by manual annotations) were already masked by Word2Vec. This highlights how PETRE self-adjusts to the starting anonymization, and minimizes the number of masked terms needed to reach the desired privacy threshold.

A more accurate picture of the actual utility preserved by the anonymized outcomes can be gathered by analyzing the NMI metric. Results reported in Figure 2e suggest that the effect of anonymization on document clustering is not that dramatic, with the lowest NMI being 66%. Moreover, they indicate that PETRE has a small impact on empirical utility, with an NMI similar to that of the starting anonymization. Indeed, this difference (5%) is smaller than the difference between the average NMI for NER-based and PPDP-oriented approaches (10%).

The mean runtimes of PETRE for each document are reported in Table 2. From these figures, the SHAP explainability module took more than 90% of the time, which is in line with the discussion in Section 5. Because the documents and the privacy threshold were the same for all starting anonymizations, differences in runtimes are only determined by the number of

iterations required to obtain the desired privacy level. Since our method masks one term per iteration, the longer the runtime, the more terms had to be masked. The numbers are thus directly proportional to the recall/privacy improvement rate of the different methods, being much larger for the (privacy lacking) NER-based methods than for PPDP-grounded approaches. Figure 4 illustrates this, exhibiting a near perfect positive correlation ($r > 0.99$, $p\text{-}value < 0.00001$) inverse to that observed in Figure 3.

Table 2: Mean PETRE runtime per document for each starting anonymization

Starting anonymization	Runtime (seconds)
Stanford NER3	54.57
Stanford NER4	31.89
Stanford NER7	52.18
Presidio	36.88
Word2Vec	7.48
k-anonymity random	9.65
k-anonymity greedy	9.43

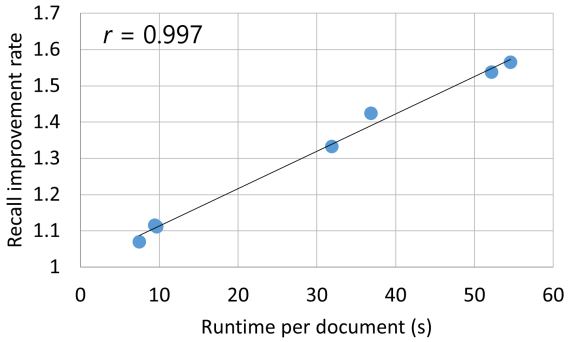


Figure 4: Correlation between PETRE's runtime and its recall improvement rate

6.6.2. By k values

To analyze the influence of the privacy parameter, we conducted experiments with values of $k \in \{2, 3, 5, 7, 10\}$. Figure 5 depicts the mean PETRE runtime per document for each value of k and starting anonymization.

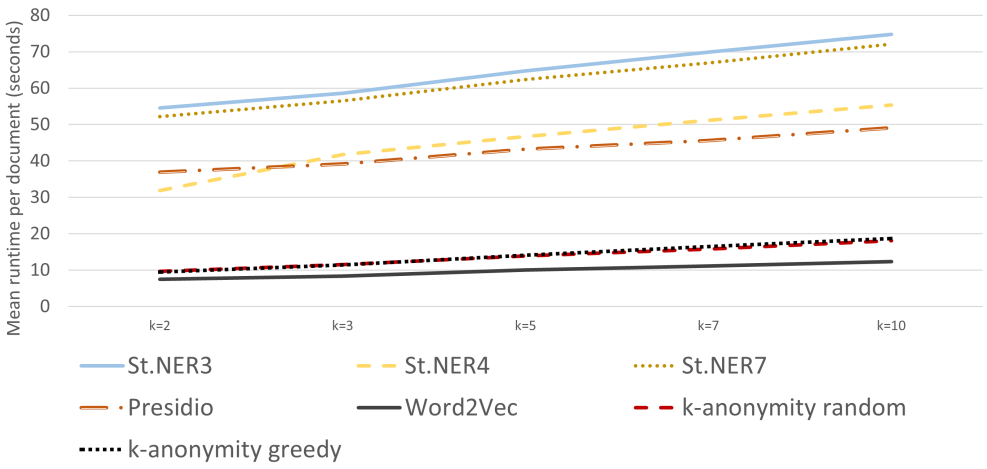


Figure 5: PETRE mean runtime per document for different k values and starting anonymizations

As initially noted for $k = 2$ in Table 2, NER-based approaches consistently demanded more processing time (due to the need to mask more terms and, therefore, more PETRE iterations) than PPDP-oriented methods to meet the same k-anonymity level. Most of the runtime for NER-based methods was spent to achieve $k = 2$, which indicates that the majority of missed terms were those enabling unequivocal re-identification (i.e, $k = 1$). These are either identifiers or quasi-identifiers that did not belong to the predefined categories supported by the employed NER model. In contrast, PPDP-oriented methods expended, in average, a similar amount of time (8-10 seconds) to reach $k = 2$ from the starting anonymization as they did to reach $k = 10$ from $k = 2$. This illustrates that these methods missed a similar amount of identifiers and unique sets of quasi-identifiers as low-frequency (i.e., $k < 10$) quasi-identifiers.

As discussed in Section 6.5, the manual anonymizations employed as groundtruth barely reached a rank of $k = 2$. This invalidates meaningful precision and recall assessments for $k > 2$, as most of the additional terms masked by PETRE to reach those larger k would be considered mismatches with the (much weaker) manual anonymization. Nevertheless, to further analyze the practical privacy enhancement brought by PETRE with respect to k, Figure 6 reports the mean rank of the PETRE-enhanced outcomes for each value of k and starting anonymization.

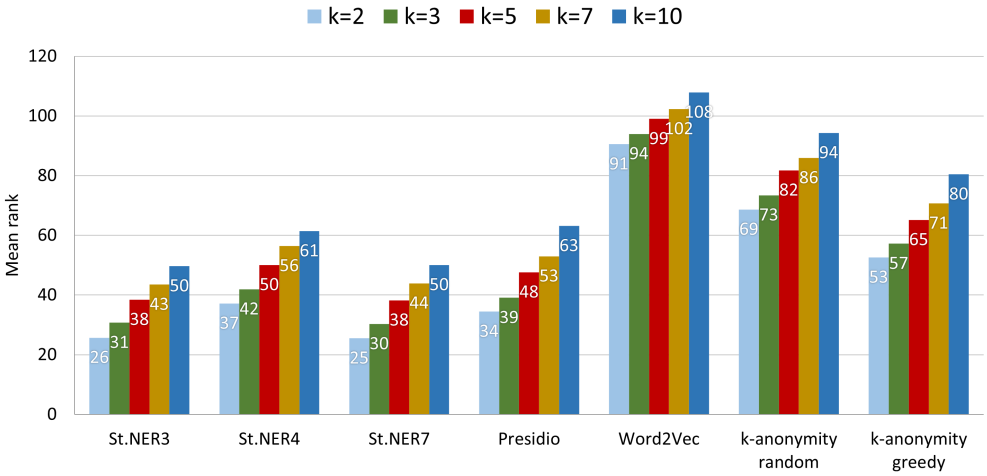


Figure 6: Mean ranks for each PETRE-enhanced starting anonymization with different values of k

It is noteworthy that rank values are significantly larger than the requested k. This is the result of starting documents that were already non-identifiable by the re-identification model, resulting in nearly random distributions with a rank for the correct identity in the order of hundreds, most of the times. The very large magnitude of these outlying ranks, even if they only happen in a small set of document, significantly increases the rank average. Also, and especially for larger values of k, it is quite common that even a single masking could result in an increase in the observed rank larger than the privacy threshold. Regardless of the absolute value, it is more relevant to observe the proportional rank increase as k becomes larger. Also, consistently with the recall values reported in the previous section, we observe that the mean ranks obtained for $k = 2$ (and subsequent values) are significantly larger for PPDP-oriented methods than for NER-based approaches.

6.6.3. By background knowledge

As stated in Section 4.1, background knowledge can greatly influence the re-identification accuracy. This, in turn, would affect PETRE anonymization, resulting in masking more or less terms for a given k. So far, we have considered a worst-case scenario in which the background knowledge that can be leveraged to conduct re-identification attacks consists in the very

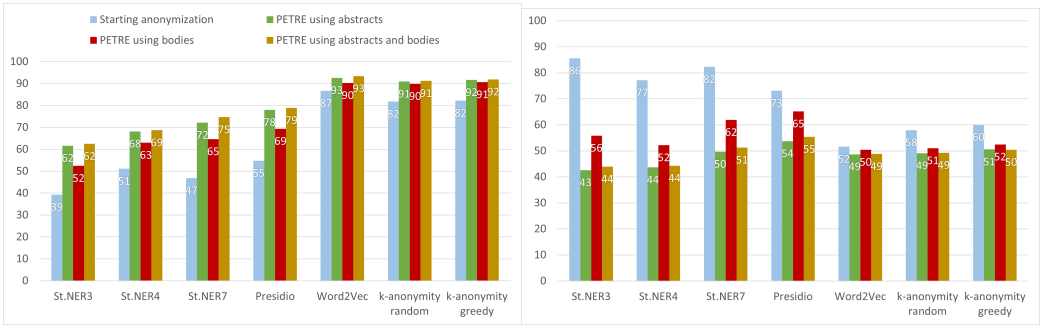
same abstracts to be anonymized. As discussed in Section 6.3 this setting seems to better match with how manual anonymization was conducted. To evaluate the influence of the background knowledge in PETRE's behavior, in this section, we define two additional knowledge sets.

The first one includes the bodies of the Wikipedia articles instead of the abstracts. Because bodies and abstracts partially overlap, this can be viewed as a more realistic scenario, in which attackers have partial (rather than complete) knowledge of the data to be anonymized. Because bodies are significantly longer than abstracts, training the re-identification model took 2 hours and 22 minutes. For this scenario, we should consider that 8% of the articles did not have a body, thereby resulting in no additional privacy enhancement because it was not possible to re-identify them with no background knowledge available.

The second set includes both the bodies and the abstracts. Even though bodies do not add any new information that would be subjected to protection (because all this information was already present in the abstracts to be anonymized), they may be beneficial for building the re-identification model. In particular, because the re-identification model is built by fine tuning a general-purpose language model (DistilBERT), the additional in-domain text provided by the bodies may contribute to better adapt the baseline language model to the task and domain of application. Being this the largest set, the training of the re-identification model took 3 hours and 32 minutes.

Figure 7 reports the evaluation results for these two new sets and compares them with those already reported in Figure 2.

The impact of PETRE (increase of recall and decrease of precision and NMI) is diminished when employing only the bodies, and slightly reinforced when using both the abstracts and the bodies. In the first case, the impact reduction can be explained by the terms appearing in the abstracts but not in the bodies, which would not be accounted as (quasi-)identifiers anymore, as they do not belong to the background knowledge. All these terms would not be subjected to masking, thereby resulting in a less strict, but more utility-preserving anonymization. The small set of articles that did not have bodies to be used as background knowledge also contributed to reduce PETRE's impact, as no masking was done for those. Differences are more noticeable for NER-based methods, which suggests that they tend to miss more abstract-exclusive terms. There are grounds to believe that these terms correspond to unique sets of potential quasi-identifiers (in case they would be available in the background knowledge set). First, they are neglected



(a) Recall

(b) Precision

(c) F_1-score

(d) F_2-score

(e) NMI

Figure 7: Evaluation metrics of starting anonymizations and their PETRE-anonymized counterparts for different background knowledge sets

by the manual anonymization, since not masking them increases precision more than decreases recall; and, as argued in Section 6.5, human annotators may overlook complex combinations of quasi-identifiers. Secondly, they are mostly spotted by PPDP-oriented methods, as recall and precision are barely

affected for the different knowledge sets; and these methods are specifically designed to detect unique combinations of terms across the dataset to be protected, regardless of their availability in public sources.

As expected, using both the bodies and the abstracts results in minimal differences over using only the abstracts. As discussed above, adding the bodies should not increase the amount of terms that need to be masked, as all terms in the abstracts already appear in the background knowledge. But, at the same time, the additional corpus (i.e., the bodies) leveraged to train the re-identification model can contribute to make it sightly more accurate; and this higher accuracy requires few more maskings to reach a desired k, as shown by the minor but consistent increase of recall across methods.

6.6.4. Generalizability to domain-specific data

We next test the ability of PETRE to generalize to the legal-specific documents in the TAB dataset with Standford NER7 and Presidio, which were the two NER-based methods that resulted in the best privacy-utility balance in the previous experiments. Table 3 reports the evaluation figures for both starting anonymizations and their PETRE-enhanced versions for $k \in \{2, 3, 5, 7, 10\}$.

Table 3: Results for St.NER7 and Presidio with PETRE in the TAB dataset

Anonymization	Recall	Precision	F_1-score	F_2-score	NMI
St.NER7	40.39	54.85	43.24	41.19	60.13
St.NER7+PETRE k=2	81.10	44.20	57.99	63.51	52.97
St.NER7+PETRE k=3	82.20	41.37	56.87	62.67	36.72
St.NER7+PETRE k=5	82.71	38.98	55.34	61.44	33.66
St.NER7+PETRE k=7	83.17	36.78	54.41	60.67	35.37
St.NER7+PETRE k=10	83.65	34.64	53.01	59.49	35.65
Presidio	70.46	72.28	70.92	71.46	63.28
Presidio+PETRE k=2	79.31	58.22	67.87	71.17	61.96
Presidio+PETRE k=3	80.84	55.27	66.56	70.44	56.86
Presidio+PETRE k=5	81.68	52.08	64.74	69.21	58.63
Presidio+PETRE k=7	82.49	48.22	63.69	68.49	38.52
Presidio+PETRE k=10	82.93	46.09	62.01	67.24	37.16

Similarly to what was observed in Figure 2, applying PETRE with k=2 results in a significant privacy improvement (recall) while retaining most

of the empirical utility (NMI). This effect is (again) especially noticeable for starting anonymizations offering limited protection, as it is the case of NER7. In fact, both starting anonymizations converge to similar recall values when enhanced with PETRE, which suggests consistent masking: regardless of the starting anonymization, PETRE ends up masking the same set of terms. We can thus conclude that PETRE effectively generalizes to domain-specific data, showing a similar behavior to that observed on the general-purpose dataset.

For $k > 2$, as discussed in Section 6.6.2, the fact that the manual groundtruth barely reached a $k = 2$ rank largely invalidates precision and recall analysis (and, by extension, F-scores). For those values of k, we better rely on the empirical utility assessment provided by NMI. As expected, increasing k decreases NMI by up to 25%. For NER7, most of this decrease happens at $k = 3$, while for Presidio, significant utility degradation does not occur until $k = 7$. This disparity suggests that NER7 introduces more unnecessary masking, making it more sensitive to PETRE's protection in terms of NMI impact.

7. Conclusions and future work

We presented PETRE, a risk-driven automatic method for enhancing the privacy of any text anonymization. Its application is more straightforward than the other PPDP-oriented methods proposed in the literature while reaching competitive results. The user's involvement is limited to defining the background knowledge, which can be omitted if a worst case scenario is considered, and the privacy threshold k. Afterwards, both the re-identification model training and PETRE's iterative masking operate in an unsupervised and self-controlled way. On the contrary to NER-based methods, masking can be tailored to each individual to be protected according to what information is included in the background knowledge. Also, the desired level of protection can be intuitively defined in terms of maximum residual re-identification risk. Being our method grounded in probabilistic k-anonymity, this level of protection is attained regardless of the data to be protected, so no *a posteriori* empirical evaluations of risk would be needed. To sum up, we empower practitioners with an intuitive and flexible method that allows to automatically assess and minimize the residual re-identification risk of their data.

Experiments show that PETRE significantly enhances the privacy protection of NER-based anonymizations, and that results are consistent regardless of the starting anonymization method. Our empirical risk assessment also reveals substantial deficiencies in the protection attained by manual annotations. We posit that this deficiency is attributable to inherent human limitations in detecting complex combinations of quasi-identifiers, which can be exploited by machine learning models to enable re-identification.

The iterative nature of PETRE, in combination with the computational cost of SHAP [25] (currently the costlier operation), make runtime the main drawback of the approach. As future work, we intend to leverage alternative explainability methods with lower explanation runtimes. In addition, we plan to examine the types of terms that the explanations indicate facilitate or impede re-identification (i.e., positive or negative weights). This analysis seeks to achieve a more nuanced understanding of the factors impacting the re-identification. Future research could also explore the replacement of masked terms rather than their suppression. This will help better preserve the semantics/utility of the protected documents. For this purpose, it would be worth investigating the usage of optimization algorithms, such as [11, 49], to find the set of textual replacements that best preserves utility while still meeting the privacy requirement.

Acknowledgments

Partial support to this work has been received from the EU's Horizon 2020 research and innovation program under grant agreement no. 871042 "SoBigData++", the Norwegian Research Council ("CLEANUP" project, grant nr. 308904), the MCIN/AEI/ 10.13039/501100011033 and ERDF "A way of making Europe" under grant PID2021-123637NB-I00 "CURLING", the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez and grant 2021SGR-00115), and INCIBE and European Union NextGenerationEU/-PRTR (project "HERMES" and INCIBE-URV Cybersecurity Chair).

References

[1] Abril, D., Navarro-Arribas, G., Torra, V., 2012. Improving record linkage with supervised learning for disclosure risk assessment. Information Fusion 13, 274–284.

[2] Agency, E.M., 2014. European medicines agency policy on publication of clinical data for medicinal products for human use (policy/0070; ema/240810/2013).

[3] Arthur, D., Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, USA. pp. 1027–1035.

[4] Bier, E., Chow, R., Golle, P., King, T.H., Staddon, J., 2009. The rules of redaction: Identify, protect, review (and repeat). IEEE Security & Privacy 7, 46–53.

[5] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S.M., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y., 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR abs/2303.12712. URL: https://doi.org/10.48550/arXiv.2303.12712, doi:10.48550/ARXIV.2303.12712, arXiv:2303.12712.

[6] Canada, H., 2019. Guidance document on public release of clinical information. URL: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.

[7] Cornell Law School, 2007. Rule 49.1 privacy protection for filings made with the court. URL: https://www.law.cornell.edu/rules/frcrmp/rule_49.1.

[8] Csányi, G.M., Nagy, D., Vági, R., Vadász, J.P., Orosz, T., 2021. Challenges and open problems of legal document anonymization. Symmetry 13, 1490.

[9] 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.

[10] Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A., 2021. The limits of differential privacy (and its misuse in data release and machine learning). Communications of the ACM 64, 33–35.

[11] Dorigo, M., Stützle, T., 2019. Ant colony optimization: overview and recent advances. Springer.

[12] Dwork, C., 2006. Differential privacy, in: International Colloquium on Automata, Languages and Programming, Springer. pp. 1–12.

[13] El Emam, K., 2013. Guide to the de-identification of personal health information. CRC Press.

[14] European Commission, 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[15] Fernandes, N., Dras, M., McIver, A., 2019. Generalised differential privacy for text document processing, in: International Conference on Principles of Security and Trust, Springer. pp. 123–148.

[16] Ferrández, O., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M., 2012. Evaluating current automatic de-identification methods with veteran's health administration clinical documents. BMC medical research methodology 12, 1–16.

[17] Fred, A.L.N., Jain, A.K., 2003. Robust data clustering, in: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA, IEEE Computer Society. pp. 128–136. URL: https://doi.org/10.1109/CVPR.2003.1211462, doi:10.1109/CVPR.2003.1211462.

[18] Hassan, F., Domingo-Ferrer, J., Soria-Comas, J., 2018. Anonymization of unstructured data via named-entity recognition, in: International Conference on Modeling Decisions for Artificial Intelligence, Springer, Mallorca, Spain. pp. 296–305.

[19] Hassan, F., Sánchez, D., Domingo-Ferrer, J., 2023. Utility-preserving privacy protection of textual documents via word embeddings. IEEE Transactions on Knowledge and Data Engineering 35, 1058–1071.

[20] Huang, Y., Song, Z., Chen, D., Li, K., Arora, S., 2020. Texthide: Tackling data privacy in language understanding tasks, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Punta Cana, Dominican Republic. pp. 1368–1382.

[21] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P., 2012. Statistical Disclosure Control. volume 2. Wiley New York, New York.

[22] Johnson, A.E., Bulgarelli, L., Pollard, T.J., 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers, in: Proceedings of the ACM Conference on Health, Inference, and Learning, Association for Computing Machinery, Toronto, Ontario, Canada. pp. 214–221.

[23] Lan, G., Hu, M., Li, Y., Zhang, Y., 2023. Contrastive knowledge integrated graph neural networks for chinese medical text classification. Engineering Applications of Artificial Intelligence 122, 106057. URL: https://www.sciencedirect.com/science/article/pii/S0952197623002415, doi:https://doi.org/10.1016/j.engappai.2023.106057.

[24] Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L., 2021. Anonymisation models for text data: State of the art, challenges and future directions, 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. pp. 4188–4203.

[25] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Association for Computing Machinery. pp. 4768–4777.

[26] Mamede, N., Baptista, J., Dias, F., 2016. Automated anonymization of text documents, in: IEEE Congress on Evolutionary Computation, IEEE, Vancouver, BC, Canada. pp. 1287–1294.

[27] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D., 2014. The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland, USA. pp. 55–60.

[28] Manzanares-Salor, B., Sánchez, D., Lison, P., 2024. Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. Data Mining and Knowledge Discovery URL: https://link.springer.com/article/10.1007/s10618-024-01066-3, doi:https://doi.org/10.1007/s10618-024-01066-3.

[29] Mendels, O., 2020. Nlp approaches to data anonymization. URL: https://towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929.

[30] Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H., 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology 10, 1–16.

[31] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781 .

[32] Mosallanezhad, A., Beigi, G., Liu, H., 2019. Deep reinforcement learning-based text anonymization against private-attribute inference, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Hong Kong, China. pp. 2360–2369.

[33] Neamatullah, I., Douglass, M.M., Lehman, L.W.H., Reisner, A., Villarroel, M., Long, W.J., Szolovits, P., Moody, G.B., Mark, R.G., Clifford, G.D., 2008. Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making 8, 1–17.

[34] Nyffenegger, A., Stürmer, M., Niklaus, J., 2023. Anonymity at risk? assessing re-identification capabilities of large language models. CoRR

abs/2308.11103. URL: https://doi.org/10.48550/arXiv.2308.11103, doi:10.48550/ARXIV.2308.11103, arXiv:2308.11103.

[35] Papadopoulou, A., Lison, P., Øvrelid, L., Pilán, I., 2022. Bootstrapping text anonymization models with distant supervision, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 4477–4487.

[36] Patsakis, C., Lykousas, N., 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. Scientific Reports 13, 16026.

[37] Pilán, I., Lison, P., Ørelid, L., Papadopoulou, A., Sánchez, D., Batet, M., 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. Computational Linguistics 48, 1053–1101.

[38] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should i trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 1135–1144.

[39] Samarati, P., 2001. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13, 1010–1027.

[40] Sánchez, D., Batet, M., 2016. C-sanitized: A privacy model for document redaction and sanitization. Journal of the Association for Information Science and Technology 67, 148–163.

[41] Sánchez, D., Batet, M., 2017. Toward sensitive document release with privacy guarantees. Engineering Applications of Artificial Intelligence 59, 23–34.

[42] Sánchez, D., Batet, M., Viejo, A., 2013. Automatic general-purpose sanitization of textual documents. IEEE Transactions on Information Forensics and Security 8, 853–862.

[43] Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108 .

[44] Singhal, S., Zambrano, A.F., Pankiewicz, M., Liu, X., Porter, C., Baker, R.S., 2024. De-identifying student personally identifying information with gpt-4, in: Proceedings of the 17th International Conference on Educational Data Mining, pp. 559–565.

[45] Soria-Comas, J., Domingo-Ferrer, J., 2012. Probabilistic k-anonymity through microaggregation and data swapping, in: 2012 IEEE International Conference on Fuzzy Systems, pp. 1–8.

[46] Staddon, J., Golle, P., Zimny, B., 2007. Web-based inference detection, in: USENIX Security Symposium, Association for Computing Machinery, Boston, MA, USA. pp. 1–16.

[47] Subakti, A., Murfi, H., Hariadi, N., 2022. The performance of bert as data representation of text clustering. Journal of Big Data 9, 1–21.

[48] U.S. Department of Labor, 2004. The health insurance portability and accountability act. URL: `https://www.dol.gov/agencies/ebsa/laws-and-regulations/laws/hipaa`.

[49] Wang, D., Tan, D., Liu, L., 2018. Particle swarm optimization algorithm: an overview. Soft computing 22, 387–408.

[50] Yang, H., Garibaldi, J.M., 2015. Automatic detection of protected health information from clinic narratives. Journal of Biomedical Informatics 58, S30–S38.

[51] Yogarajan, V., Mayo, M., Pfahringer, B., 2018. A survey of automatic de-identification of longitudinal clinical narratives. arXiv:1810.06765 .