Neural Text Sanitization with Explicit Measures of Privacy Risk

Anthi Papadopoulou
Language Technology Group, University of Oslo
anthip@ifi.uio.no

Yunhao Yu
École Polytechnique
yunhao.yu@polytechnique.edu

Pierre Lison
Norwegian Computing Center
plison@nr.no

Lilja Øvrelid
Language Technology Group, University of Oslo
liljao@ifi.uio.no

Abstract

We present a novel approach for text sanitization, which is the task of editing a document to mask all (direct and indirect) personal identifiers and thereby conceal the identity of the individuals(s) mentioned in the text. In contrast to previous work, the approach relies on explicit measures of privacy risk, making it possible to explicitly control the trade-off between privacy protection and data utility.

The approach proceeds in three steps. A neural, privacy-enhanced entity recognizer is first employed to detect and classify potential personal identifiers. We then determine which entities, or combination of entities, are likely to pose a re-identification risk through a range of privacy risk assessment measures. We present three such measures of privacy risk, respectively based on (1) span probabilities derived from a BERT language model, (2) web search queries and (3) a classifier trained on labelled data. Finally, a linear optimization solver decides which entities to mask to minimize the semantic loss while simultaneously ensuring that the estimated privacy risk remains under a given threshold. We evaluate the approach both in the absence and presence of manually annotated data. Our results highlight the potential of the approach, as well as issues specific types of personal data can introduce to the process.

1 Introduction

Personal data, also known as Personally Identifiable Information (PII), often abound in text documents, from emails to patient records, court judgments, interview transcripts or customer service chats. Protecting the privacy of the individuals mentioned in those documents is an important task, particularly for sensitive texts which might disclose confidential information such as health status, religious beliefs, ethnicity or sex life.

It is, however, possible to apply privacy-enhancing techniques such as *text sanitization* to conceal the identity of those individuals from the texts, and thereby make it easier to share data to third parties, in particular for the purpose of scientific research or statistical analysis. The goal of text sanitization is to transform a document through edit operations such as hiding particular text spans or replacing them by more general values. Although complete anonymization compliant with data privacy frameworks such as the General Data Protection Regulation (GDPR, 2016) has been shown to be very difficult to achieve in practice (Weitzenboeck et al., 2022), text sanitization can substantially enhance the level of privacy protection while simultaneously retaining most of the semantic content expressed in the documents.

Existing work on text sanitization has primarily focused on masking predefined entity types through sequence labelling (Dernoncourt et al., 2017; Liu et al., 2017; Jensen et al., 2021). These previous approaches, however, may not mask enough PII to prevent re-identification, as they are restricted to a fixed list of semantic categories to detect. These are often named entities such as persons, organizations, or locations. As a consequence, personal information that do not belong to those predefined categories (for instance, mentions of a person's appearance or occupation) will be ignored. Paradoxically, they may also end up masking *too much* information, as they systematically mask all occurrences of a given entity type (for instance, all locations) regardless of the actual influence of a particular entity on the risk of re-identifying the individuals mentioned in the original document (Lison et al., 2021).

In this paper we present a novel approach to text sanitization that seeks to address these limitations. The approach relies on a privacy-enhanced entity recognizer that goes beyond named entities and can detect demographic attributes and other types of personal information that frequently occur in text. The integration of empirical measures of privacy

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 217–229
November 20–23, 2022.

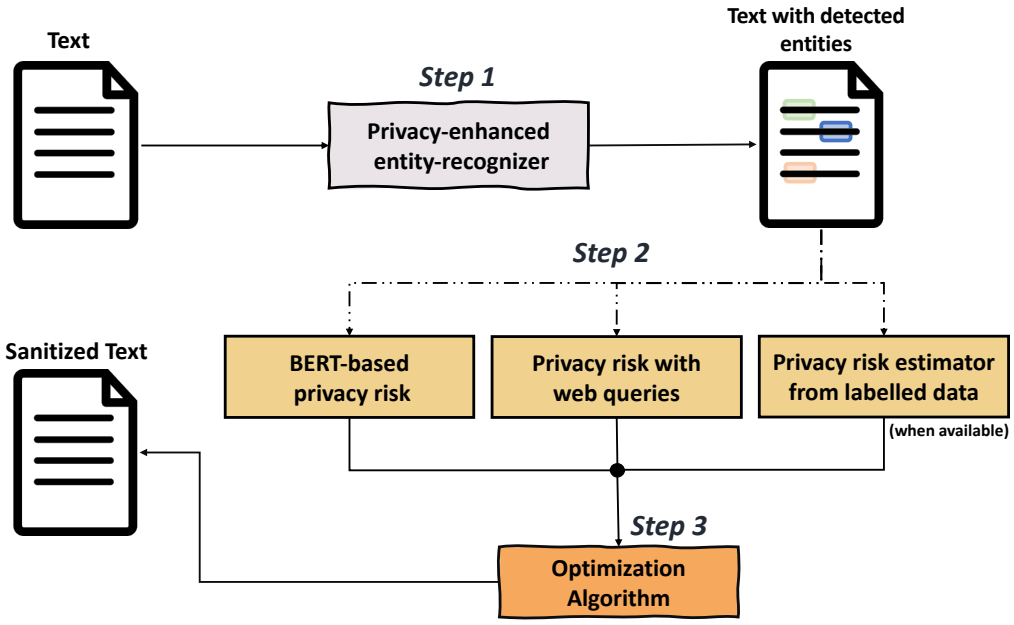

Figure 1: General sketch of the approach. The text document is first given as input to the privacy-enhanced entity recognizer which detects personal information present in the text, along with their semantic type. Then three privacy risk measures are used to determine which entities may constitute a privacy risk. Finally, an optimization algorithm makes the optimal masking decisions for each document, resulting in a sanitized text.

risk also makes it possible to strike an explicit balance between data utility and privacy protection. The resulting risk measures are fed to an optimization solver which determines the optimal set of entities to mask in each document. Figure 1 provides a general outline of the procedure. The code along with the models used is publicly available.[1]

The proposed approach can be applied without any labelled data, provided there already exists a generic Named Entity Recognizer (NER) and a version of Wikidata for the language employed in the documents. If text annotated with masking decisions is available, the approach can take advantage of them to further enhance the model's performance. The modularity of the approach also allows for the integration of additional methods to measure the privacy risk associated with the entities mentioned in the text.

This paper makes the following contributions:

- A neural entity recognizer specifically tailored for privacy protection, based on the combination of a generic NER model with a gazetteer derived from Wikidata.
- Several methods for empirically estimating the re-identification risk associated with the presence of a given entity or combination of entities in a document. One method relies on probabilities derived from BERT, while a second relies on web search queries, and a third one on a neural classifier trained from labelled data (when available).
- A pipeline that combines the neural entity recognizer with privacy risk measures and an optimization algorithm to determine the optimal set of entities to mask, given a privacy risk threshold and estimates of semantic loss.
- Evaluation results based on the recently developed Text Anonymization Benchmark (Pilán et al., 2022) that demonstrate the validity of the approach both in the absence and presence of in-domain labelled data.

The structure of the rest of the paper is the following. A background and review of related work are provided in Section 2. Section 3 details our approach, followed by an evaluation and discussion in Section 4. We conclude in Section 5.

Terminological note

The removal of PII from text documents to protect the identity of the individuals mentioned in those

[1] https://github.com/NorskRegnesentral/NeuralTextSanitizer

texts has received multiple names in the literature, such as de-identification, pseudonymization, sanitization and anonymization (Deleger et al., 2013; Eder et al., 2019; Sánchez and Batet, 2016; Lison et al., 2021). Following (Sánchez and Batet, 2016; Brown et al., 2022), we settle in this paper on the term "sanitization" to differentiate it from techniques traditionally termed as "de-identification" (Dernoncourt et al., 2017; Yogarajan et al., 2018), which are restricted to specific semantic categories. Moreover we wish to avoid the use of the term "anonymization", as it is notoriously difficult to precisely define what qualifies as anonymous data in relation to legal frameworks such as GDPR (Hintze, 2017), particularly when it comes to unstructured data (Weitzenboeck et al., 2022).

2 Background

Privacy is a fundamental human right, and various legal frameworks for data protection[2] have been put in place in recent years to ensure that individuals remain in control of their personal data. Those frameworks specify strict guidelines on how data that may contain personal information should be collected, stored and processed. Personal identifiers can be divided in two broad categories (Elliot et al., 2016; Domingo-Ferrer et al., 2016):

Direct identifiers: Information that can irrevocably and uniquely identify an individual (e.g. name, social security number, email address, bio-metric data, etc.)

Quasi identifiers: Information that cannot directly single out an individual, but may do so indirectly when combined with other quasi identifiers (e.g. date of birth, occupation, city of residence, ethnicity etc.). For instance, the combination of gender, date of birth and postal code can single out between 63 and 87% of the U.S. population (Golle, 2006).

Both direct and quasi identifiers need to be masked (i.e. removed or generalized) to prevent identity disclosure. This necessarily leads to a a loss of information or data utility, and the objective of text sanitization is therefore to determine the set of masking operations that ensure the privacy risk remains below a given threshold, yet preserve as much data utility as possible.

[2]See e.g. the General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA) in the US or China's Personal Information Protection Law (PIPL).

NLP approaches to text sanitization have mostly focused on medical data, using either rule-based methods (Ruch et al., 2000; Douglass et al., 2005) or sequence labelling models trained on manually annotated data for pre-defined categories (Deleger et al., 2013; Dernoncourt et al., 2017; Liu et al., 2017; Johnson et al., 2020).

Text sanitization approaches have also been developed in the field of privacy-preserving data publishing (PPDP). Those approaches seek to enforce a privacy model by searching for the optimal set of masking decisions to ensure that the requirements of the model are met. The k-anonymity privacy model (Samarati and Sweeney, 1998) has been adapted for text data in k-safety (Chakaravarthy et al., 2008) and k-confusability (Cumby and Ghani, 2011). Like k-anonymity, these approaches require every entity to be indistinguishable from k-1 other entities. t-plausibility (Anandan et al., 2012) is a similar model which depends on PII being already detected to perform generalization so as to ensure that at least t documents can be derived through specialization of the generalized terms. Finally C-sanitized (Sánchez and Batet, 2016) is designed to mimic human annotators by taking into account semantic inferences in the text, in addition to disclosure risk. To this end, mutual information scores are calculated manually from co-occurrence counts in web data. Those PPDP approaches, however, typically treat the text simply as a flat collection of terms, missing thus the importance of context for the entities and the linguistic inter-relationships between these terms.

Pilán et al. (2022) present the Text Anonymization Benchmark (TAB), a corpus of court judgements from the European Court of Human Rights (ECHR), manually enriched with detailed annotations on the PII expressed in each document. The authors also propose a set of novel evaluation metrics for the task as well as baseline results using a neural sequence labelling model. Papadopoulou et al. (2022) describe a bootstrapping approach for text sanitization based on k-anonymity. Their approach requires, however, an explicit specification of the background knowledge associated with each individual, which may be difficult to acquire.

The masking operations employed in text sanitization are non-perturbative (i.e. limited to either hiding text spans or replacing them by more general values). This need to preserve the "truth value" of the original document is important for

Category	Explanation	Examples
CODE	flight numbers, case ids, passport numbers	3086/23, LH3042
ORG	companies, schools, hospitals	Budapest Police Department, Ministry of Justice
DATETIME	dates, time, duration of event	23 November 2006, 7, 12 and 5 months
LOC	city names, addresses	Austria, Martin County
QUANTITY	money values, percentage of a value	6,932 Ukrainian hryvnyas, two
PERSON	names, nicknames, translations	Joe Smith, The Rock
DEM	nationality, occupation, education	artist, Italian, MSc in Astrophysics
MISC	vehicles, tools, process	aircraft, gun, liquidation

Table 1: Categories of semantic types along with some selected subcategories and examples taken from the silver corpus.

many types of data releases: a clinical report in which the description of symptoms and diagnosis has been randomly altered would be of little interest for e.g. medical researchers. This requirement distinguishes text sanitization from other privacy-enhancing methods based on differential privacy (Feyisetan et al., 2019; Krishna et al., 2021), which transform existing texts through the addition of artificial noise. Although those techniques are undeniably useful to create texts (or text representations) that can enforce specific privacy guarantees, they address a different task than the one discussed in this paper, as they effectively produce new, synthetic texts instead of masked versions of existing documents (Pilán et al., 2022).

3 Approach

In the following we introduce the three steps of our neural text sanitization model.

3.1 Privacy-enhanced entity recognizer

Accurately detecting all potential PII in a text is a crucial first step in a text sanitization approach, since it ensures that subsequent steps will have potentially sensitive text spans available while arriving at the necessary masking decisions.

Generic NER systems are commonly used as part of anonymization solutions such as Microsoft's Presidio[3]. Such systems, however, often fail to detect demographic attributes (e.g. occupation, sexual orientation, medical condition) or other miscellaneous information (e.g. tools, vehicles, field of work, or manner of death) that are potential quasi-identifiers.

To address this limitation, we combine a generic NER model with a gazetteer including terms typically employed as attributes of human individuals in Wikidata. More specifically, we inspected 3646 Wikidata properties related to humans and manually identified those that could potentially belong to either DEM (demographic attributes associated to a person, such as their profession, ethnicity or family status) or MISC (any other information that may contribute to identifying a person, but is not an "attribute" of that person). We end up with 44 DEM properties and 196 MISC properties, which we used to create the gazetteer. Some examples of four Wikidata properties filtered as DEM and MISC respectively are:

- *occupation* (P106) –> writer, builder, professor etc.
- *political ideology* (P1141) –> progressivism, democrat, antimilitarism etc.
- *cause of death* (P509) –> nitric acid poisoning, suicide, helicopter crash etc.
- *convicted of* (P1399) –> forgery, matricide, home invasion etc.

The combination of the generic NER model with this gazetteer allows us to recognize a total of 8 categories of PII, detailed in Table 1.

To further enhance the performance of the entity recognition (and counteract the limited coverage of the gazetteer), we then apply the NER model and the gazetteer to create a *silver corpus* of PII. Our training data consists of 2500 Wikipedia summaries and 2500 ECHR cases as they are publicly and freely available sources of data that are rich in PII. This silver corpus is then employed to fine-tune a neural language model – more specifically RoBERTa (Liu et al., 2019) to label text spans according to the 8 categories in Table 1.

We split the silver corpus into a training (90%), development(10%), and test dataset(10%). The average text length in the silver corpus is 14 sentences, keeping in mind that ECHR cases are typically longer documents than Wikipedia biographies.

[3]https://github.com/microsoft/presidio

Figure 2 shows the distribution of semantic types of the silver corpus for the three dataset splits.

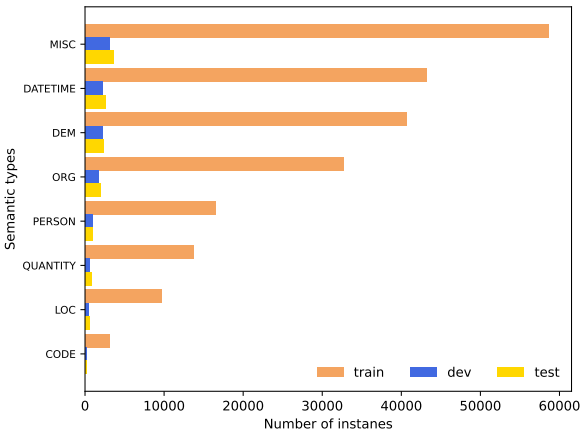

Figure 2: Distribution of semantic types on the train, development and test split of the silver corpus of PII

While manually inspecting some of the training instances we also notice examples of label confusion which can be attributed to Wikidata. Some property values, which are entered by editors for each Wikidata page, belonged to the wrong semantic type (e.g. dates, organization names or nationalities in properties such as cause of death). We thus expect to see some examples of these types of errors by the model.

3.2 Privacy risk measures

Once text spans expressing potential PII are detected in the document, the next step is to determine the privacy risk associated with their presence in the document. Indeed, not all of the entities detected in the previous step will need to be masked. To determine the entities, or combinations of entities, that constitute a re-identification risk and need to be masked, we rely on several complementary measures, detailed below.

3.2.1 Language model probabilities

One heuristic to automatically determine whether an entity or a combination of entities need to be masked is to use a language model to calculate surprisal measures in the form of the probability of the text span in its document context. Intuitively, a more "surprising" entity corresponds to a PII with a larger information content, and therefore a higher re-identification risk. Conversely, a text span that can be predicted from the rest of the document will typically correspond to information that is less specifically tied to the individual to protect.

We use a pre-trained RoBERTa model with a language modeling head on top (linear layer) to calculate the log probability of each text span detected by the privacy-enhanced entity recognizer. In case the span consists of more than one token, we compute the final probability by adding the log probabilities of each token. A span with a low log-probability corresponds to an entity that is difficult to predict and thus more informative/specific. A threshold is then established to determine which entities need to be masked on the basis of those log-probabilities. In practice, this threshold can be selected empirically.

3.2.2 Privacy risks with web queries

The re-identification risk can also be estimated using web queries. Intuitively, the idea is to query a web search engine with a particular combination of entities, and check whether web results also mention the person to protect, in which case the entities pose an unacceptable re-identification risk and need to be masked. For instance, if we wish to conceal the mention of Annalena Baerbock from a document, the combination of the two entities "Germany" and "minister" will correspond to a privacy risk, as the search for those words on Google yields among the top results web pages that do mention the name of Annalena Baerbock.

To avoid the need to crawl web pages to search for the mention of the person to protect, we start by querying the search engine for the person name, and store the results. This makes it possible to find out whether a combination of entities is dangerous by computing the intersection of the URLs related to the person and the URLs related to the entities. If this intersection is non-empty, at least one web search result contains both the person name and the combination of entities. Due to practical constraints with web search APIs, the algorithm only extracts the top k results for each search query. Our implementation currently relies on Google as search engine and a value of k set to 50.[4]

Admittedly, sending queries to a search engine is costly, since a document may comprise hundreds of entities, and querying a web search engine with their various combinations is a time-consuming process. To address this issue, we also emulate the results obtained by Algorithm 1 using a neural model. More specifically, the model seeks to predict whether a combination of entities is likely to

[4]The search results were gathered in June 2022. Search results might differ depending on when they were acquired.

```
def find_risky_entity_combinations
    (entities, person_name, max_arity):
# entities: text spans detected in document
# person_name: name of individual to protect
# max_arity: max size of combined entities to query

  # (Initially empty) set of entity combinations
  # that can re-identify the person
  risky_entity_combs ← ∅

  # We search the person on the web
  urls_for_person ← search(person_name)

  # We start by searching for single entities,
  # then pairs of entities, up to max arity
  for n = 1 → max_arity:

    # We loop on all entity combinations of size n
    for entity_comb in combine(entities, n):

      # We search the entities (joined by "AND")
      urls_for_entities ← search(entity_comb)

      # We also augment the URLs about the person
      urls_for_person ← urls_for_person
      + search(person_name + entity_comb)

      # If at least one URL is in both sets, those
      # entities can lead to re-identification
      if urls_for_entities ∩ urls_for_person ≠ ∅:
        Add entity_comb to risky_entity_combs

  return risky_entity_combs
```

Algorithm 1: Procedure for determining which entities, or combination of entities, can uncover the identity of the person to protect, based on web search queries.

lead to web search results that mention the person name. The neural model employed for this prediction task relies on contextualized embeddings from BERT, together with an LSTM layer to compute a single embedding vector for each entity. The model is trained on the search results for 20 documents in the training set of the TAB corpus. See the Appendix for details on the architecture.

3.2.3 Classifier trained on labelled data

Finally, one can also measure the privacy risk associated with entities mentioned in a text through a supervised model. More specifically, one can collect text documents manually annotated by human experts with masking decisions and train a neural model to reproduce those masking decisions.

Our implementation relies on a fine-tuned RoBERTa neural language model that takes as input a text including the occurrences of each entity in its document context and the semantic category produced by Step 1. The language model is augmented with a classification head (after pooling), and is fined-tuned on the labelled data to predict whether a given entity should be masked.

3.3 Optimization algorithm

The privacy risk measures described in the previous sections generates a list of entities, or combinations of entities, that constitute an unacceptable re-identification risk. When single entities are marked as risky, the corresponding decision is trivial: the entity must be masked. However, risky *combinations* of entities are more difficult to handle, as we need to decide on which subset of entities to mask or possibly retain in clear text.

We formulate this task as a linear programming problem[5] where the objective is to minimize the semantic loss subject to the constraint that, for each combination of entities deemed risky, at least one entity in the combination must be masked. The semantic loss is then defined as the sum of the information content IC for all masked entities. This semantic loss is a measure of quantifying the information lost when entities are masked, i.e. the usability of the resulting text if certain PII is missing. Formally, the optimization problem is defined as:

$$\text{Minimize} \quad \sum_{e \in E_d} \mathit{masked}(e)\ \mathit{IC}(e)$$

subject to the constraints:

$$\sum_{e \in \mathit{ent\_tuple}} \mathit{masked}(e) \geq 1$$

$$\forall\ \mathit{ent\_tuple} \in \mathit{risky\_entity\_combinations}_d$$

where:

- E_d is the set of entities detected by the privacy-enhanced entity recognizer for document d
- *masked(e)* is a binary variable that takes a value of 1 if the entity e is masked and 0 otherwise
- $IC(e)$ is the information content of entity e, defined as the negative log-probability of e according to BERT, as done in Section 3.2.1. If the entity contains several words, the log-probabilities of each word are summed.
- $\mathit{risky\_entity\_combinations}_d$ is the list of all entity combinations detected in document d by the entity recognizer and categorized as risky by at least one privacy risk measure.

[5]The CP-SAT Solver from Google OR-tools was used in our implementation.

4 Evaluation

We evaluate the proposed approach on the Text Anonymization Benchmark (TAB) (Pilán et al., 2022) which consists of 1268 ECHR court judgements manually annotated for text anonymization benchmarking. Court judgements are freely available documents that are not subject to data protection regulations. The annotations in TAB identify all possible PII in the texts, associated with both a semantic category (e.g., person name, code, demographic property, etc.) and a masking decision.

The majority of entity types in the TAB corpus belong to the DATETIME (34.6%), ORG (26.3%), and PERSON (15.7%) semantic types, while 63.4% of all the annotations were masked as quasi identifiers and 4.4% as direct identifiers (mainly CODE and PERSON semantic types), with the rest of the detected spans being left as is in the text (Pilán et al., 2022). The test set, which we use for our evaluation purposes, consists of 127 documents which were annotated and quality checked by more than one annotators.

We first analyse the performance of the privacy-enhanced entity recognizer, and then evaluate the performance of the complete pipeline.

4.1 Entity recognition

We evaluate the privacy-enhanced entity recognition model from Section 3.1 on the test set of TAB, using the full set of manually detected PII prior to masking. We compare the performance of our system against two baselines: (i) the generic NER model used in the first step of the silver corpus creation, and (ii) the generic NER model in combination with the gazetteer populated with Wikidata properties related to human individuals. The latter comparison aims to evaluate whether the neural model fine-tuned on the silver corpus generalizes to unseen PII not included in the gazetteer. The generic NER model corresponds to a RoBERTa language model fine-tuned for named entity recognition on the Ontonotes corpus (Weischedel et al., 2011). Table 2 provides the evaluation results. See Appendix for details on training parameters.

The results show that the privacy-enhanced entity recognizer model is able to detect with reasonable accuracy almost all semantic types apart from the MISC category, for which it seems to have the lowest performance. MISC is a broad semantic type that cannot be concretely categorised, and is thus difficult for a model to predict; for instance the longer MISC example in the TAB test dataset is a quote of 49 tokens. Since MISC entities are derived from Wikidata properties, we also do not expect them to completely match the MISC entities found in the court judgments of the TAB corpus.

Below are some example of recognition errors, where the left side corresponds to a manually annotated text span as seen in the TAB corpus, while the right side corresponds to the spans detected by the entity recognizer:

- British national [DEM] - British [DEM]
- discrimination case [MISC] - discrimination [MISC]
- five attacks [QUANTITY] - five [QUANTITY] attacks [MISC]
- life imprisonment [DATETIME] - life imprisonment [MISC]
- without a father for an important part of its childhood years [MISC] - father [DEM] childhood years [MISC]

Those examples illustrate that a mismatch in the entity label or text span boundary (compared to the manually annotated texts) does not necessarily mean that the model fails to detect a PII.

4.2 Full sanitization model

We now analyse the performance of the complete pipeline (in various variants) on the task of deciding which entity to mask in a given document. We adopt the evaluation metrics put forward by (Pilán et al., 2022) to assess the performance of text sanitization methods. In particular, we provide separate recall measures for the direct and quasi identifiers, as well as both an unweighted and weighted precision score, the latter taking into account the informativeness of each span (Pilán et al., 2022).

Baselines

We compare the approach presented in this paper against three baselines:

- **Mask all entities from generic NER**: this baseline simply considers that all named entities (as detected by the neural NER model fine-tuned on Ontonotes) constitute a privacy risk and need to be masked.
- **Mask all entities from privacy-enhanced recognizer**: same as above, but with entities extracted with the privacy-enhanced recognizer from Section 3.1.

	CODE			ORG			PERSON			DATETIME			LOC			QUANTITY			DEM			MISC		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Gen. NER	.98	.79	.88	.62	.91	.74	.97	.64	.77	.90	.99	.94	.34	.92	.50	.39	.75	.51	.77	.42	.54	.03	.26	.05
Gen. NER+Gaz.	.97	.93	.95	.78	.95	.86	.98	.95	.96	.93	.94	.94	.72	.90	.80	.95	.72	.81	.28	.73	.40	.10	.36	.15
Enhanced ER	.98	.97	.97	.76	.96	.87	.98	.98	.98	.92	.99	.95	.53	.89	.66	.42	.84	.56	.27	.76	.40	.10	.32	.15

Table 2: Token-level precision, recall and F_1 score by entity type on the test set of the TAB corpus. The results include the two baselines (generic NER model, either alone or augmented with the gazetteer with terms extracted from Wikipedia properties) as well as the privacy-enhanced entity recognizer fine-tuned on the silver corpus. Labels such as ORG and LOC are considered to be interchangeable, as many entities of those types can be assigned to both, as is the case for e.g. country names.

- **Mask most specific entities**: this baseline only considers as risky the entities of type CODE, PERSON, DATETIME, LOC or QUANTITY extracted with the privacy-enhanced recognizer, which were most frequently masked in the TAB corpus. Entities of other types are not considered to constitute a privacy risk.

Privacy risk measures

As explained in 3.2.1, the BERT-based privacy risk relies on a threshold to determine whether an entity or combination of entities should be seen as a privacy risk (based on log probabilities). The threshold is selected empirically based on the development set of the TAB corpus (see Appendix), and set to a value t = -3.5. We also include in the evaluation the privacy risk measure based on web queries from Section 3.2.2 and the neural model trained on labelled data from the training section of the TAB corpus.

Table 3 provides the evaluation results, split into two distinct scenarios, a *zero-shot* scenario in the absence of manually labelled data, and a *fine-tuned* scenario where the TAB training corpus was used to both further fine-tune the privacy-enhanced entity recognizer and also train a supervised model to predict whether an entity should be masked.

For the zero-shot scenario, we can observe that the two baselines (*Generic NER, Privacy-enhanced recognizer*) tend to over-mask the text. The probabilities derived by the LM model (*BERT-based risk*) show a relatively high recall on both direct and quasi identifiers, but a lower precision score, while the opposite holds for the strategy based on risk measures from the emulated web queries.

Unsurprisingly, the performance increases when manually labelled data is available (fine-tuned scenario). The two baselines for this category (*Privacy-enhanced + FT, Mask all* and *Mask most specific*) show both a high precision and recall score, as the detected PII comes closer to the manual annotations. For the LM probabilities we notice a slight drop in precision, which is presumably due to longer spans (especially for the MISC category) which were masked by the risk measure but not the annotators. The web model on the other hand shows a higher recall score and a lower precision score. Finally, the risk measure that is best able to balance data utility and privacy risk is the classifier trained on manual data (*Supervised risk*).

We can observe from Table 3 that the weighted precision score is generally higher than the uniform precision. This indicates that the false positives were of a more general nature so their information content was low. This gives us a better overview of the utility of the masked text. An example text from the test dataset with different masking decisions can be found in the Appendix.

We conduct an error analysis on the two optimal approaches for each scenario and we notice two trends. On the one hand, the masking strategies failed to mask some entities that the annotators decided to mask (mainly dates, locations, laws, foreign words e.g. *Florida, England, 1987, CPT/Inf (2000)17, önlisans etc.*)

We also notice a trend of partial masking, which results in partial or correct masking decisions, something that is not reflected in the evaluation results as they do not match with any of the decisions made by the annotators. Some examples, where the left side corresponds to the human annotation and the right the decision made by one of the two masking strategies, are:

- United Kingdom nationals [MASK] - United Kingdom [MASK]
- medical secretary [MASK] - secretary [MASK]
- SEK 147,000 (approximately 15,800 euros [EUR]) [MASK] - SEK 147,000 [MASK] 15,800 euros [EUR] [MASK]

Entity recognition	Masking strategy	P	WP	R_{all}	R_{direct}	R_{quasi}	F_1
		Zero-Shot					
Generic NER	Mask all	.41	.58	.91	.95	.88	.57
Privacy-enhanced	Mask all	.44	.52	.96	.99	.94	.60
Privacy-enhanced	BERT-based risk	.57	.62	.91	.98	.83	.70
Privacy-enhanced	Web query risk	.82	.84	.50	.66	.40	.62
Privacy-enhanced	BERT-based risk + Web query risk	**.57**	**.60**	**.91**	**.99**	**.84**	**.70**
		Fine-tuned					
Privacy-enhanced + FT	Mask all	.52	.57	.98	.99	.97	.68
Privacy-enhanced + FT	Mask most specific	.76	.77	.84	.98	.87	.83
Privacy-enhanced + FT	BERT-based risk	.54	.58	.95	.99	.89	.69
Privacy-enhanced + FT	Web query risk	.64	.68	.84	.91	.78	.73
Privacy-enhanced + FT	Supervised risk	**.79**	**.81**	**.89**	**.99**	**.89**	**.84**
Privacy-enhanced + FT	Supervised risk + Web query risk	.64	.69	.94	.99	.93	.76
Privacy-enhanced + FT	All three risk measures	.54	.58	.97	.99	.95	.69

Table 3: Evaluation results on the test portion of the TAB corpus.
- "Privacy-enhanced": privacy-enhanced entity recognizer from Section 3.1
- 'Privacy-enhanced + FT": same model after fine-tuning on the semantic labels from the TAB training set.
- "BERT-based risk": masking strategy in which text spans indicated as risky by the BERT-based risk measures (Section 3.2.1), using the optimization algorithm from Section 3.3 to make the final decisions.
- "Web based risk": similar strategy, this time using the results from emulated web queries as risk measures.
- "Mask most specific": mask the entities of type CODE, PERSON, DATETIME, LOC or QUANTITY.
- "Supervised risk" refers to the risk measure based on a neural model estimated from the masking decisions of human experts in the training set of the TAB corpus.

P=Precision, *WP*=Weighted precision, as defined in (Pilán et al., 2022), R_{all}=Recall for all identifiers, R_{direct} = Recall for direct identifiers, R_{quasi} = Recall for quasi identifiers (as annotated in the TAB corpus), and F_1 = harmonic mean of precision and recall on all identifiers. The best results are highlighted in bold.

- 25 April, 24 May, 16 June, 6 July and again on 27 July 1994 [MASK] - 25 April [MASK] 24 May [MASK] 16 June [MASK] 6 July [MASK] 27 July 1994 [MASK]

The task of text sanitization can have many different but correct masking solutions, as long as the identity of the individual is protected. Evaluating against one gold standard is very useful since we can judge the extend of the usefullness of the approaches we propose. However, it also means that the evaluation is limited by the (sometimes subjective) decisions made by the annotators.

5 Conclusion

This paper presented a novel approach to automated text sanitization. The approach relies on the detection of different types of PII as well as empirical measures of re-identification risk based on language models, web queries, and (when available) manually labelled data. Such an approach makes it possible to derive explicit estimates of the privacy risk associated with a given masked document. Those estimates can be employed to find the most appropriate trade-off between data utility and privacy protection, depending on the particular requirements of the application.

The approach is evaluated on the newly released Text Anonymization Benchmark (Pilán et al., 2022). The evaluation results demonstrate the potential of the approach – both in the presence and absence of manually labelled data –, but also highlight the difficulty of the task.

Future work will focus on refining the privacy-enhanced entity recognizer, to improve the detection of MISC entities. We also aim to investigate more flexible masking strategies, such as the replacement of detected entities by more general text spans (such as [Orléans] being replaced by [city in France]), instead of merely hiding the entities from the text. Finally, we wish to explore evaluation measures that do not rely on manually labelled data, as text sanitization is a task that may admit several, equally valid solutions (Lison et al., 2021).

6 Acknowledgements

We acknowledge support from the Research Council of Norway (CLEANUP project, grant nr. 308904).

References

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. 2012. T-plausibility: Generalizing words to desensitize text. *Trans. Data Privacy*, 5(3):505–534.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. 2008. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 843–852. ACM.

Chad M. Cumby and Rayid Ghani. 2011. A machine learning based system for semi-automatically redacting documents. In *IAAI*.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. 2016. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Synthesis Lectures on Information Security, Privacy & Trust. Morgan & Claypool Publishers.

M.M. Douglass, G.D. Cliffford, A. Reisner, W.J. Long, G.B. Moody, and R.G. Mark. 2005. De-identification algorithm for free-text nursing notes. In *Computers in Cardiology*, pages 331–334.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269, Varna, Bulgaria. INCOMA Ltd.

Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. 2016. *The Anonymisation Decision-Making Framework*. UKAN.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

GDPR. 2016. General Data Protection Regulation. European Union Regulation 2016/679.

Philippe Golle. 2006. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM Workshop on Privacy in electronic society*, pages 77–80. ACM.

Mike Hintze. 2017. Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency. *International Data Privacy Law*, 8(1):86–101.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference of Computational Linguistics (NODALIDA)*.

Alistair EW Johnson, Lucas Bulgarelli, and Tom J Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 214–221.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *J. of Biomedical Informatics*, 75(S):S34–S42.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization.

P. Ruch, R. H. Baud, A. M. Rassinoux, P. Bouillon, and G. Robert. 2000. Medical document anonymization with a semantic lexicon. *Proceedings of the AMIA Symposium*, pages 729–733.

Pierangela Samarati and Latanya Sweeney. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Emily Weitzenboeck, Pierre Lison, Malgorzata Agnieszka Cyndecka, and Malcolm Langford. 2022. The GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*.

Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. 2018. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv preprint arXiv:1810.06765*.

A Appendix

Privacy-enhanced entity recognizer

Table 4 details the parameters used to train the privacy-enhanced entity recognizer described in Section 3.

Parameter	
Optimizer	AdamW
Learning rate	2e-5
Loss function	CrossEntropy
Inference layer	Linear
Epochs	3
Full fine-tuning	yes
GPU	yes
Early stopping	yes

Table 4: Training Parameters for the RoBERTa model

BERT-based privacy risk

Figure 3 shows an example of a precision-recall curve used to determining thresholds for the BERT-based privacy risk measure. We calculated a general precision and recall score for different thresholds and chose one that shows a good balance between privacy risk and data utility. Stricter thresholds favor recall but result in a low precision score, while more lenient thresholds showed a drop in recall but better precision score.

Neural model emulating web queries

The architecture described in Section 3.2.2 is presented below in Figure 4.

Example of masking decisions

We also present in Figure 5 an example of different masking decisions (see for a text from the TAB test dataset, as mentioned in Section 4.2.

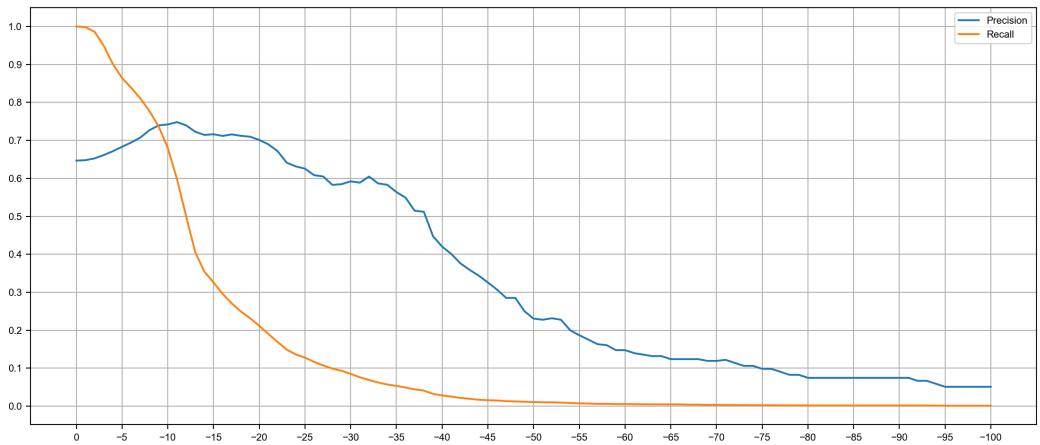

Figure 3: Precision-Recall curve for determining appropriate thresholds

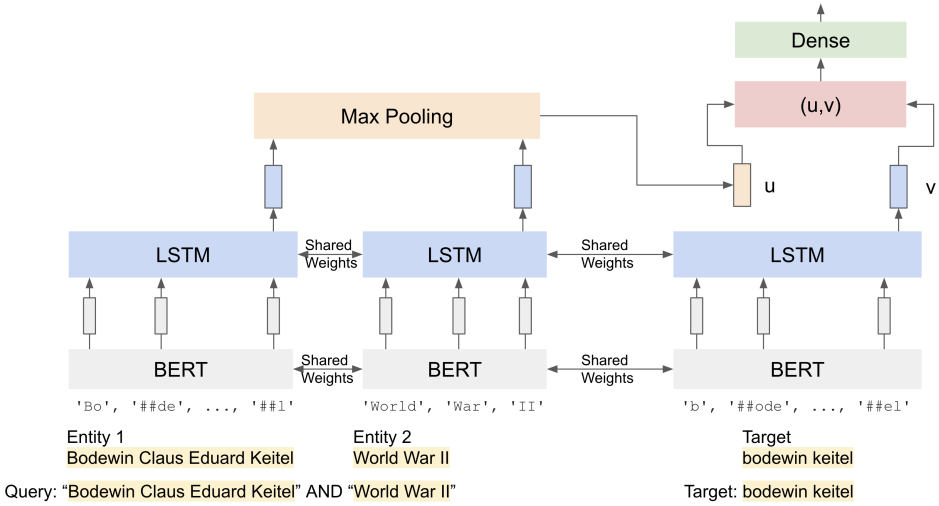

Figure 4: Architecture of the web query model

The case originated in an application (no. 27961/02) against the United Kingdom of Great Britain and Northern Ireland lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a British national, Mr Tony Booth ("the applicant"), on 25 October 2001. The applicant was represented by Royds Rdw, solicitors in London. The United Kingdom Government ("the Government") were represented by their Agent, Mr C. Whomersley of the Foreign and Commonwealth Office, London. The applicant complained under Articles 8 and 14 of the Convention and Article 1 of Protocol No. 1 that, because he was a man, he was denied social security benefits equivalent to those received by widows. On 17 November 2005 the Court decided to communicate the complaints concerning widows' benefits.

The applicant was born in 1944 and lives in Sussex. His wife died on 29 October 2000. They had no children from the marriage. His claim for widows' benefits was made on 2 January 2001 and was rejected on 31 May 2001 on the ground that he was not entitled to widows' benefits because he was not a woman. The applicant did not appeal as he considered or was advised that such a remedy would be bound to fail since no such social security benefits were payable to widowers under United Kingdom law.

Figure 5: Example of masking decisions on the excerpt of an ECHR court case. The *blue line* denotes masking decisions made by a human annotator. The *grey line* corresponds to text spans to be masked after being detected by the privacy enhanced entity-recognizer and passed through the two privacy risk measures. Finally, the *orange line* shows spans to be masked after detection by the fine-tuned entity-recogniser (fine-tuned on the TAB training dataset) and the three risk assessments mentioned in Table 3.