Bootstrapping Text Anonymization Models with Distant Supervision

Anonymous ACL submission

Abstract

We propose a novel method to bootstrap text anonymization models based on distant supervision. Instead of requiring manually labeled training data, the approach relies on a knowledge graph expressing the background information assumed to be publicly available about various individuals. This knowledge graph is employed to automatically annotate text documents including personal data about a subset of those individuals. More precisely, the method determines which text spans ought to be masked in order to guarantee k-anonymity, assuming an adversary with access to both the text documents and the background information expressed in the knowledge graph. The resulting collection of labeled documents is then used as training data to fine-tune a pre-trained language model for text anonymization. We illustrate this approach using a knowledge graph extracted from Wikidata and short biographical texts from Wikipedia. Evaluation results with a BERT-based model and a manually annotated collection of 553 summaries showcase the potential of the approach, but also unveil a number of issues that may arise the knowledge graph is noisy or incomplete. The results also illustrate that, contrary to most sequence labeling problems, the text anonymization task may admit several alternative solutions.

1 Introduction

Personal data is ubiquitous in text documents. Due to this presence of personal information, many text sources fall under the scope of data protection regulations such as GDPR (GDPR, 2016). As a consequence, they cannot be shared with third parties (or even used for other purposes than the one originally intended when collecting the data) without a proper legal ground, such as the explicit consent of the individuals to whom the data refers.

In case obtaining the consent of all those individuals is unfeasible, an alternative is to *anonymize* the data to ensure those individuals can no longer be identified. Anonymization is often defined as the complete and irreversible process of removing all Personally Identifiable Information (PII) from a dataset (Elliot et al., 2016a). Such PII includes both direct identifiers such as person names, passport numbers or mobile phone numbers, but also more indirect information such as date of birth, gender, nationality or workplace that can also lead to (re-)identification when combined with one another (Domingo-Ferrer et al., 2016).

The anonymization of text data is, however, a difficult challenge for which many open questions remain (Lison et al., 2021). One important problem is the lack of labeled corpora for this task, making it difficult to train data-driven text anonymization models in many domains. The few datasets that currently exist mainly focus on the medical domain (Dernoncourt et al., 2016; Bråthen et al., 2021) and are typically limited to predefined categories of entities[1]. Models trained on such datasets are also known to be difficult to transfer to new domains (Johnson et al., 2020; Hartman et al., 2020).

We present in this paper an alternative approach for training text anonymization models. Crucially, this approach does not require access to manually labeled training data. Rather, we adopt a distant supervision approach that revolves around a *knowledge graph* expressing the background information assumed to publicly known on various individuals. The approach proceeds in three steps:

1. The knowledge graph is first converted into an inverted index, making it possible to efficiently compute the set of individuals associated with a given combination of entities.

[1] This task of detecting and masking predefined semantic categories (such as names, organizations and locations) is often called *de-identification*. In contrast, text *anonymization* is not limited to a fixed set of semantic categories, but must consider how any textual element may influence the risk of disclosing the identity of the person referred to in the text (Chevrier et al., 2019; Lison et al., 2021).

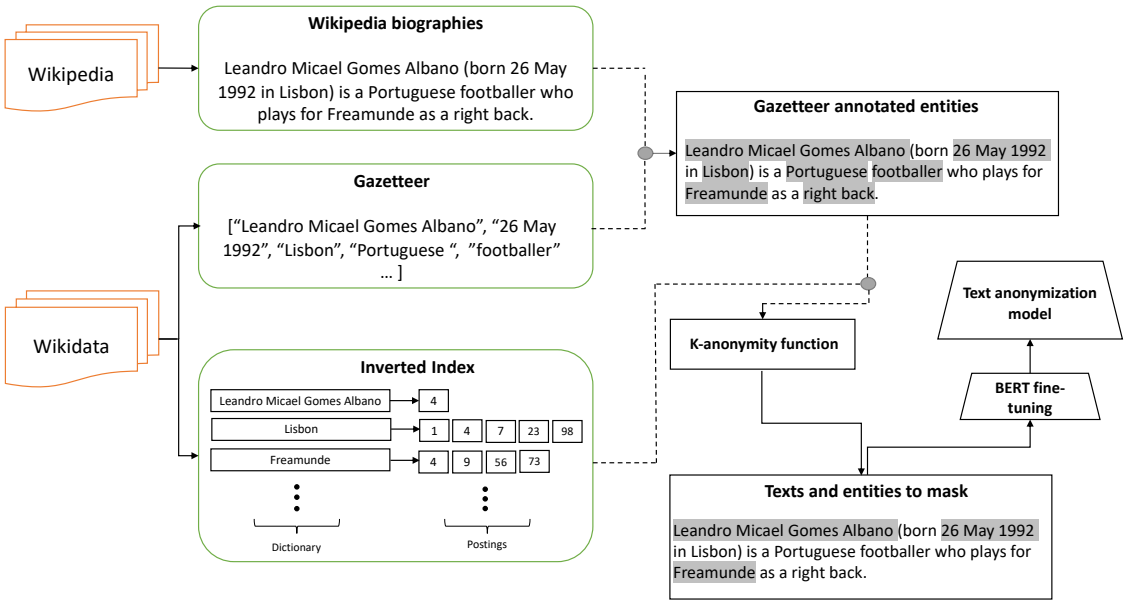

Figure 1: General sketch of the approach, illustrated with some examples for clarity

2. The inverted index is then employed as distant supervision source (Mintz et al., 2009; Liang et al., 2020) to automatically annotate a collection of text documents including personal data. The goal of this annotation process is to determine which tokens to mask in order to guarantee k-anonymity (that is, to guarantee that the information conveyed in the anonymized document is sufficiently general to be shared by at least k individuals).

3. Finally, this labeled collection of documents is used as training data to fine-tune a large, pre-trained language model (in our case BERT) for the task of determining which text span to mask in a given document.

The proposed approach has several benefits. As it relies on distant supervision, there is no need for manually annotating text documents with text spans to mask, a procedure that is costly and time-consuming. The approach also follows a *privacy-first* strategy that determines which terms to mask based on a privacy model (k-anonymity). This strategy provides an explicit account of the disclosure risk associated with a given set of masking decisions on a document, using the knowledge graph to represent the information that can be drawn upon by an adversary to uncover the identity of the individual(s) we seek to protect. This account of disclosure risk makes it possible to adjust the trade-off between data protection and data utility. Finally, the approach makes it arguably easier to port text anonymization models to new languages and domains, as knowledge graphs can often be reused across multiple languages and text genres.

The validity of the approach is evaluated through experiments with a collection of short biographical texts extracted from Wikipedia. Wikipedia biographies constitute an ideal test-bed for the proposed approach, as these texts contain a lot of PII, including both direct and quasi identifiers. Those biographical texts were then automatically annotated with text spans to masks using a knowledge graph derived from Wikidata. The general procedure is illustrated in Figure 1.

This paper makes three main contributions:

- A novel, privacy-first approach to the training of data-driven text anonymization models in the absence of labeled data.
- An implementation of that approach with a large knowledge graph derived from Wikidata, which is applied to automatically label biographical texts from Wikipedia with text spans to mask.
- A new dataset of 553 Wikipedia summaries manually annotated for sensitive information, which we use to evaluate the empirical performance of the proposed approach.

The rest of the paper is structured as follows. The next section reviews related work on text

anonymization. Section 3 describes the three steps of our approach, which is then evaluated in Section 4. We conclude in Section 5.

2 Related Work

As stipulated by Article 8 of the European Convention on Human Rights and Article 12 of the Universal Declaration of Human Rights, privacy is a fundamental right and an essential component of a democratic society. To ensure that every person remains in control over their own personal data, legal frameworks such as the General Data Protection Regulation (GDPR) (GDPR, 2016) in the European Union spell out how personal data should be collected, processed and shared.

Personally identifiable information can be divided into main categories (Elliot et al., 2016b):

Direct identifiers: information that can be used to directly single out an individual, such as the person name, social security number, email or physical address, passport number, bio-metric records, mobile phone number, etc..

Quasi identifiers: information that is not univocally related to a unique individual, but may nevertheless lead to re-identification when combined with other quasi identifiers[2] such as date of birth, gender, ethnicity, religion, employer, city of residence, etc.

Although most existing work on anonymization focuses on quantitative tabular data, several studies have also investigated the problem of anonymizing text data, either from an NLP perspective or from the field of data privacy, and in particular privacy-preserving data publishing (PPDP).

The first NLP approaches relied on rule-based methods for pattern detection mainly in medical text documents (Douglass et al., 2005). Recent approaches on the anonymization of medical health records focus on detecting direct identifiers and quasi identifiers using sequence labelling models trained from manually annotated data (Deleger et al., 2013; Dernoncourt et al., 2016; Liu et al., 2017; Hathurusinghe et al., 2021).

One drawback of these NLP approaches is that they are typically limited to detecting predefined (semantic) categories of identifiers and quasi identifiers, without taking into account other types of information that may uncover the identity of the person. For instance, the physical appearance of a person or their professional activities will often provide clues about the person identity, yet rarely belong to the semantic categories to detect. In addition, those methods typically mask all detected text spans uniformly, without making it possible to parametrize the anonymization process based on the estimated disclosure risk.

PPDP approaches to text anonymization, on the other hand, generally seek to enforce a *privacy model* such as k-anonymity (Samarati and Sweeney, 1998), then search for the optimal set of masking operations – such as removal or generalization of the original values – to ensure the privacy model requirements are met.

The k-anonymity model was adapted for text data in the k-safety and k-confusability models (Chakaravarthy et al., 2008; Cumby and Ghani, 2011). Both approaches require every sensitive entity to be indistinguishable from at least $k-1$ other entities. The entities are then generalized to become indistinguishable and thus, safe from disclosure risk. The t-plausibility model (Anandan et al., 2012) introduced a similar approach based on the generalization of (already detected) terms, seeking to ensure that at least t documents can be derived through specialization of the generalized terms. A final model is C-sanitize (Sánchez and Batet, 2015) which provides a priori privacy guarantees by relying on the mutual information of the sensitive entities and the rest of the words in the document. The words that are more likely to lead to identification of the sensitive term to be protected, either individually or in combination with others, are then generalized. The mutual information scores used in C-sanitize are derived from co-occurrence counts in web data.

PPDP approaches makes it possible to explicitly adjust the trade-off between data protection and data utility. However, many PPDP approaches rely on the assumption that sensitive entities are already detected in a preprocessing step. They also often rely on external resources that may be difficult to gather (Lison et al., 2021)

Finally, recent work has investigated the use of differential privacy (Dwork and Roth, 2014) to generate synthetic texts (Sasada et al., 2021) or obfuscate documents to protect them against authorship attribution (Fernandes et al., 2019; Feyisetan et al., 2019). However, those methods operate by intro-

[2]For instance, the combination of gender, birth date and postal code has been shown to single out between 63 and 87% of the U.S. population (Golle, 2006).

ducing artificial noise either in the text or in the word representations derived from it. Contrary to the NLP and PPDP methods detailed above, those methods do not preserve the "truth value" of the document, and seek therefore to address a slightly different task than text anonymization.

3 Approach

In the following subsections, we present the three main components of our approach.

3.1 Step 1: Modeling of background information

The term *background information* refers to an attacker's possible additional knowledge that could be used to re-identify an individual in a dataset. A convenient way to express this background information is through a knowledge graph connecting individuals to protect with their various personal identifiers. This knowledge graph can be extracted from a variety of sources, such as structured databases, social network data or co-occurrence counts on web data (Sánchez and Batet, 2015).

However, knowledge graphs do not provide any efficient mechanism for determining the number of individuals associated with a particular combination of (quasi-)identifiers. This is particular problematic for quasi-identifiers that may be shared by a large set of individuals (for instance the fact that a person is male or female). To this end, we construct an *inverted index*[3] from the knowledge graph. In our case, the inverted index associates terms to individuals (or more precisely, unique indices of each individual) associated with this term. Figure 1 includes an example of inverted index where the individual with index=4 is connected with the terms "Leandro Micael Gomes Albano", "Lisbon" and "Freamunde".

Based on this inverted index, one can then efficiently query the data structure to determine the list of individuals that are related to a given set of terms. This query can be implemented through a Boolean retrieval model, taking advantage of the fact that the postings are already sorted to compute their intersection. If the resulting set set is a singleton, this means that the combination of terms allows us to uniquely re-identify the person. This is for instance the case for the combination of terms "Lisbon" and "Freamunde" in Figure 1.

An important benefit of using an inverted index to capture the relation between individuals and their quasi-identifiers is the fact that the inverted index can be easily extended to incorporate variations of a given identifier. For instance, dates and person names can be expressed in multiple formats, common nouns may have synonyms, and even locations may have alternative written variants, such as Lisbon vs. Lisboa.

3.2 Step 2: Text Anonymization with Distant Supervision

Using documents related to individuals present in this knowledge graph, we can then automatically determine which terms to mask through queries on the inverted index. The first step is to search for term occurrences in the text using a gazetteer, as illustrated in Figure 1.

Only some of the terms located by the gazetteer will need to be masked. We rely on the k-anonymity privacy model to account for the disclosure risk associated with a given set of terms in a document. k-anonymity was first introduced by Samarati and Sweeney (1998) and requires every sensitive entity to be indistinguishable from at least $k-1$ other entities based on their attributes. Through k-anonymity, the individuals can be 'hidden' by being part of a larger group. The value of k can vary depending on the dataset that needs protecting, but it should be larger than 1, since k=*1* means no anonymity. A common recommendation is to use k=*5* (Emam and Dankar, 2008), which we follow in our experiments.

Algorithm 1 is employed to determine the terms to mask in a document based on the posting lists. The algorithm starts (lines 11-14) by checking whether some terms need to be directly masked (as their presence would break k-anonymity). This is for instance the case for the term "Leandro Micael Gomes Albano", which is related to a single individual. The procedure continues by forming gradually more complex combinations of terms, and computing the intersection of their posting lists (lines 27-29). Intersections of size $< k$ represent a breach of k-anonymity, and imply that at least one of their terms must be masked. Several strategies can be followed to determine which term is most useful to mask in each combination. In this work, two strategies have been implemented. The

[3] An inverted index is a data structure commonly used in information retrieval, and consists of an index mapping terms to the documents they occurred in (Manning, 2008). Those documents are represented through a sorted list of indices, making it possible to efficiently compute intersections.

```
1  def getTermsToMask(terms, postings, maxArity,
2        termSelection, k):
3    # terms: set of terms found in a document
4    # postings: inverted index
5    # maxArity: maximum arity of the term combinations
6    # termSelect: greedySelect or randomSelect (see below)
7    # k: k-anonymity value to satisfy
8
9    termsToMask = ∅
10
11   # We mask terms associated with < k individuals
12   for term in terms:
13     if len(postings[term]) < k:
14       termsToMask ← termsToMask + term
15
16   while True:
17
18     # We create a set of possible term combinations,
19     # starting with pairs, then triples, etc.
20     termsTuples → ∅
21     for arity in [2,...maxArity]:
22       newTuples ← combine(terms - termsToMask, arity)
23       termTuples ← termTuples + newTuples
24
25     # For each term combination, we check whether the
26     # intersection of postings gives < k individuals
27     for term_1,...term_n in termTuples:
28
29       if 1 ≤ len(⋃_{i=1}^{n} postings[term_i]) < k:
30
31         # If yes, we select a term to mask
32         selectedTerm ← termSelect(term_1,..., term_n, postings)
33         termsToMask ← termsToMask + selectedTerm
34
35         # and restart the evaluation of term combinations
36         break
37
38     else: # stop when all combinations satisfy k-anonymity
39       break
40     if terms = termsToMask:  # or if all terms are masked
41       break
42
43   return termsToMask
44
45 def greedySelect(terms, postings):
46   # greedy selection: select term with shortest posting list
47   return arg min_{term_i ∈ terms} postings[term_i]
48
49 def randomSelect(terms, postings):
50   return select random term from terms
```

Algorithm 1: Extraction of terms to mask in a document, based on k-anonymity and posting lists mapping each possible term to the list of persons associated with it. When a combination of quasi-identifiers breaks the k-anonymity constraint, we either select the term with the shortest posting list in the combination (greedySelect), or choose a random term (randomSelect).

greedy strategy (lines 45-47) consists in systematically masking the most specific term – that is, the term with the shorted posting list. Alternatively, one can also select at random the term to mask in each combination.

3.3 Step 3: Fine-tuning

The two steps above result in an automatically annotated dataset that can be directly used to fine-tune a language model. Crucially, this also increases the ability of the model to generalize to texts and individuals not covered in the knowledge graph.

We frame the problem of text anonymization as a token-level sequence classification task. In this paper, we rely more specifically on BERT (Devlin et al., 2019), which is a large, transformer-based language model employed in many sequence classification tasks in the field of NLP, including recent work on data privacy (Alsentzer et al., 2019). As in most distant/weak supervision frameworks (Mintz et al., 2009; Ratner et al., 2017), the training of a generic, neural model allows us to process arbitrary texts without depending on the availability of external resources such as knowledge graphs.

4 Evaluation

The proposed approach is evaluated on short biographical texts extracted from Wikipedia, using graph data from Wikidata to determine the terms to mask to ensure $k-$anonymity. We first present the document collection and knowledge base, and then describe a manually annotated test set of biographies employed to assess the performance of the fine-tuned BERT models. We then present our results and discuss them.

4.1 Distant labelling of Wikipedia articles

The relevant background knowledge for this task comes from Wikidata [4]. Wikidata's main goal is to provide high-quality, structured data which are acquired and maintained collaboratively. It is at times used directly by Wikipedia, but typically restricted to the creation of the page's infobox.

The knowledge graph employed in this work consists of entities such as names, nicknames, translations, information on professions, dates and places of birth and death, important places, and more. To handle entities that may have several surface realizations, we augmented the inverted index to include all possible variants of a given term. This includes dates (e.g. 1992-08-05 → 5 May, 1992), person names ("Leandro Albano" → "L. Albano"), country-nationality pairs (Austrian-Austria), and alternative names for locations. A white list of very frequent terms was also established to filter out common words from the knowledge graph that are deemed generic enough not to necessitate any masking, as for example "born", "age", "man", "woman". The resulting inverted index comprises 22 034 977 terms.

This knowledge graph was then applied on a dataset of short Wikipedia biographies (Lebret

[4] https://www.wikidata.org/

et al., 2016) whose entries were filtered to consist of only humans that are also present in the knowledge graph, resulting in a total of 502 678 distinct biographies. The dataset was already split into training (80%), validation (10%), and test datasets (10%), which was preserved in this evaluation. The biographical texts are about 4 sentences long on average, with a standard deviation of 3.58.

4.2 Evaluation data

We conduct a manual annotation effort on a subset of summaries for evaluation purposes. A random sample of 553 summaries was extracted from the test dataset. The distribution of summary lengths reflects that of the test dataset, with the average being 4 sentences (11%), while around 65% were summaries with less than the average. The largest summary in the sample was 20 sentences long.

For the manual annotation, the TagTog [5] tool was used with 5 annotators, four of them undergraduate students in law, and one NLP researcher. These annotators were already familiar with the annotation task, as they had been trained and conducted similar annotation efforts in the past. They were also provided with detailed annotation guidelines and examples to follow. The objective of the annotation was to (1) find terms associated with personal information and (2) decide which of those terms ought to be masked to conceal the identity of the individual described in the biography.

Of the 553 summaries, 20 biographies were annotated by two annotators, and the rest by a single annotator. To facilitate the annotation process, the annotators were provided with pre-annotations to mark terms that were likely to express personal information. Those pre-annotations were generated by combining the gazetteer (see Section 3) with a neural NER model and a set of heuristics to recognize dates and numerical values. It should, however, be stressed that the annotators were explicitly instructed to only use those pre-annotations as a starting point and correct them as they see fit – either by modifying/deleting terms that did not include any personal information, or by inserting new terms that were ignored by the pre-annotations. See the Appendix for two annotation examples.

After this initial step of term detection, the annotators have to determine which of these terms could lead to the identification of the individual, either as direct or quasi-identifiers (see Section 2). Each term is therefore labeled as one of three mutually exclusive identifier types:

DIRECT if the term denotes a direct identifier

QUASI if the term denotes a quasi identifier

NO_MASK if the term can be left in clear text without impairing $k-$anonymity

[5] https://www.tagtog.net/

Level	Kappa	Alpha
Span	0.44	0.59
Character	0.81	0.73

Table 1: Inter-annotator agreement on the identifier type (DIRECT, QUASI OR NO_MASK).

For the 20 multi-annotated texts, we calculated inter-annotator agreement on the identifier type, by calculating Cohen's κ, as well as Krippendorff's α, with the first being based on agreement and the latter based on disagreement. These two metrics were calculated both on the span and on the character level (Artstein and Poesio, 2008) and the results are summarized for all multi-annotated documents in Table 1. It should, however, be stressed that that inter-annotator agreement for identifier types does not directly relate to the quality of the annotations, since there may be several alternative, equally correct solutions to a given anonymization task (Lison et al., 2021). This is also reinforced by the higher number of disagreement between annotators for the NO_MASK and QUASI label pairs.

4.3 Distant supervision models

We use the automatic annotations from the greedy and the random functions to train two BERT models with a linear inference layer on top (*Greedy-BERT, RandomBERT*). We used an IOB scheme to account for multi-token annotations, so that each token received either a B-MASK, I-MASK or an 'O' label. The parameters used to train the models can be found in Table 5 in the appendix.

We evaluate the performance of the models both against the automatically labeled development and test data, and on the manually annotated dataset of 553 biographies. Following the metrics proposed in the SemEval-13 task 9 (Segura-Bedmar et al., 2013), we calculate precision, recall, and F_1-score on the entity level, with two different levels of strictness, **exact** for when the boundaries of the prediction match the boundaries of the true string in an exact manner, and **partial**, when there is a

	Dev						Test					
	Precision		Recall		F1 score		Precision		Recall		F1 score	
	E	P	E	P	E	P	E	P	E	P	E	P
GreedyBERT	0.818	0.828	0.860	0.871	0.843	0.848	0.767	0.776	0.928	0.939	0.839	0.849
RandomBERT	0.804	0.815	0.876	0.882	0.838	0.847	0.723	0.732	0.969	0.974	0.828	0.835

Table 2: Entity level scores of greedy and random BERT on the automatically labeled data. We report the *exact* and *partial* results for each metric.

	Manually annotated dataset					
	Precision		Recall		F1 score	
	E	P	E	P	E	P
GreedyBERT	0.355	0.391	0.578	0.630	0.440	0.483
RandomBERT	0.352	0.387	0.605	0.661	0.445	0.488
Neural NER	0.661		0.790		0.719	

Table 3: Entity level precision, recall and F_1 score of greedy BERT, random BERT compared to the neural NER model (spaCy) on the manually labeled dataset of 553 Wikipedia summaries.

partial match between the prediction and the true string regarding boundaries.

4.4 Experimental Results

The evaluation on the automatically annotated data is meant to evaluate the feasibility of the approach, i.e., to test whether or not the background knowledge allows for a learnable annotation, whereas the manually annotated data is employed to assess the generalizability of the approach.

Table 2 shows the result for the first type of evaluation. Both models show comparable performance on the development and test datasets, and there is also a slight improvement between the exact and partial scores indicating that there are some disagreements regarding precise entity boundaries.

To contrast the annotation provided by our approach with a standard named-entity annotation task, besides the two BERT models, we also run a named entity recogniser based on the RoBERTa language model (Liu et al., 2019) and fine-tuned for NER on the Ontonotes v5 (Weischedel et al., 2011), as implemented in spaCy [6]. The performance of this system on the manually annotated dataset of summaries is shown in Table 3.

Since the manually annotated dataset also included information on the identifier type of each masked term, we also break down the recall score into recall for direct and recall for quasi identifiers, to check the performance of the models for each of these categories. Recall is the most critical metric for anonymization tasks since false negatives could directly lead to identification of the individuals we wish to protect. These results are shown in Table 4.

	Recall direct identifiers	Recall quasi identifiers
GreedyBERT	0.769	0.550
RandomBERT	0.755	0.585
Neural NER	0.775	0.755

Table 4: Entity-level recall on direct and quasi identifiers for the manually annotated dataset

4.5 Discussion

The performance of a model trained using distant supervision will necessarily depend on the quality and coverage of the knowledge base employed to generate the labels. This is also true for the approach proposed in this work, where the coverage of the knowledge graph (and of the inverted index derived from it) will influence (1) which terms will be considered as personal information and (2) which of those terms will need to be masked to enforce k-anonymity.

The experimental results illustrate some of the

[6] https://spacy.io

limitations of using Wikidata to encode the background knowledge associated with each individual. There were many instances of information mismatch between Wikipedia and Wikidata (e.g. different name spellings, information present in Wikipedia but not Wikidata). This led to either some PII not being part of the annotations or being partially annotated, which also resulted in the models often deciding to mask parts of entities instead of the entire spans, something that is reflected in the difference between exact and partial scores in Tables 2 and 3. On the other hand, the automated masking based on the inverted index also led to some spurious masking decisions, notably for terms that do not express PII but tied to a small set of individuals in Wikidata.

When testing the models on the manually annotated dataset, and comparing them against a NER system (Table 3), we see that the models' performance differs from the performance on the automatically annotated data, with spaCy's NER system outperforming them. The low precision of the models is an indication of the aforementioned issue of background information choice, since the models tend to mask information that would not generally be considered a PII, due to presence of similar terms in the inverted index.

While also analyzing the masking decisions made by the annotators we observe an over-masking trend, as well as a tendency to mask NEs more, especially for longer texts since those are the 'safer' choices (e.g. the most prominently masked categories were DEM, DATETIME, and PERSON, while regarding identifier type 56% of the masked tokens were quasi identifiers, and only 30% were left unmasked). As mentioned above, the set of masking decisions can vary a lot, which means that there is no "gold" answer, as long as the identity of the individual is protected, as also shown by the low recall on quasi identifiers in Table 4. For this reason we manually compared the output of the models, against spaCy's NER system and the manual annotations for a few texts, and we noted that despite their low scores, the two BERT models were often able to have a set of masking decisions, which despite not being similar to that of the annotator(s) or complete in the sense of entity boundaries, was able to prevent identification.

Original Text

> Jenn Mierau is a Canadian electropop musician originally from Winnipeg, who is now based in Montreal.

Human annotator

> ******* is a Canadian electropop musician originally from *******, who is now based in Montreal.

Mask from supervised NER model

> ******* is a ******* electropop musician originally from *******, who is now based in *******.

Mask from distantly supervised BERT model

> ******* ******* is a Canadian ******* originally *******, who is now based in Montreal.

The distantly supervised model produces a mask that includes the direct identifier (name), as well as including the word "from" while masking the word "Winnipeg". Despite the model's masking decisions preventing identification, this behavior is not reflected in the evaluation results against the manually annotated dataset.

5 Conclusion

We proposed a novel method to automatically annotate text documents containing personal information using background information expressed as a knowledge graph. The long-term objective of such an approach is to bootstrap text anonymization models in the absence of supervised training data, using distant supervision to determine which text spans to mask to enforce a privacy model such as k-anonymity. The automatically annotated documents can then be employed to fine-tune a pre-trained language model such as BERT.

A concrete implementation of the approach using Wikipedia biographies and Wikidata as background information is also presented. We evaluate the approach on a manually annotated set of biographies. Our experimental results demonstrate that the performance of such an approach is heavily dependent on the choice of background knowledge during implementation. The results, especially when compared to actual model output, illustrate the challenge of evaluating such a task when the acceptable pool of possible masking solutions is not limited to just one answer.

Future work will investigate several research directions. One important issue relates to how to enhance the quality of the knowledge graph, improving the coverage of quasi-identifiers while filtering out spurious terms that do not express PII. Furthermore, we aim to extend the inverted index with other sources of background knowledge beyond structured databases, and in particular co-occurrence estimates from raw, web-scale data.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Synnøve Bråthen, Wilhelm Wie, and Hercules Dalianis. 2021. Creating and evaluating a synthetic Norwegian clinical corpus for de-identification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 222–230, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K. Mohania. 2008. Efficient techniques for document sanitization. ACM Press.

Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. 2019. Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *Journal of Medical Internet Research*, 21(5):e13484.

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

L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, L. Stoutenborough, and I. Solti. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. 20(1):84–94.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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. IEEE.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. 2016a. *The anonymisation decision-making framework*. UKAN Manchester.

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

Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. 15(5):627–637.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Principles of Security and Trust - 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, Proceedings*, volume 11426 of *Lecture Notes in Computer Science*, pages 123–148. Springer.

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. ACM Press.

Tzvika Hartman, Michael D. Howell, Jeff Dean, Shlomo Hoory, Ronit Slyper, Itay Laish, Oren Gilon, Danny Vainstein, Greg Corrado, Katherine Chou, Ming Jack Po, Jutta Williams, Scott Ellis, Gavin Bee, Avinatan Hassidim, Rony Amira, Genady Beryozkin, Idan Szpektor, and Yossi Matias. 2020. Customization scenarios for de-identification of clinical notes. 20(1).

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning.

Alistair E. W. 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*, CHIL '20, page 214–221, New York, NY, USA. Association for Computing Machinery.

Remi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

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. 75:S34–S42.

Christopher Manning. 2008. *Introduction to information retrieval*. Cambridge University Press, New York.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

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

David Sánchez and Montserrat Batet. 2015. C-sanitized: A privacy model for document redaction and sanitization. 67(1):148–163.

Taisho Sasada, Masataka Kawai, Yuzo Taenaka, Doudou Fall, and Youki Kadobayashi. 2021. Differentially-private text generation via text preprocessing to reduce utility loss. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 042–047.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

R. Weischedel, E. Hovy, M. Marcus, Palmer M., R. Belvin, S. Pradhan, L. Ramshaw, and N. Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

A Appendix

Example of the setup for the annotation task mentioned in Section 4.2. Figure 2 is the result of the pre-annotation correction step (*Step 1*), and Figure 3 shows the same example but with the information one of the annotators decided to mask (*Step 2*).

Table 5 shows the parameters used to train greedy and random BERT on the automatically annotated datasets, mentioned in Section 4.3.

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

Table 5: Training Parameters for BERT models

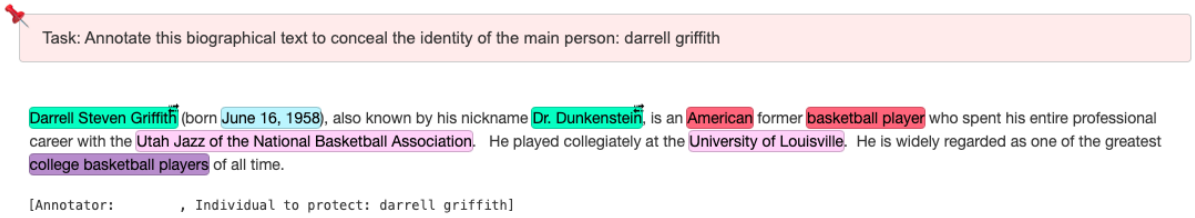

Figure 2: Step 1 of the annotation process

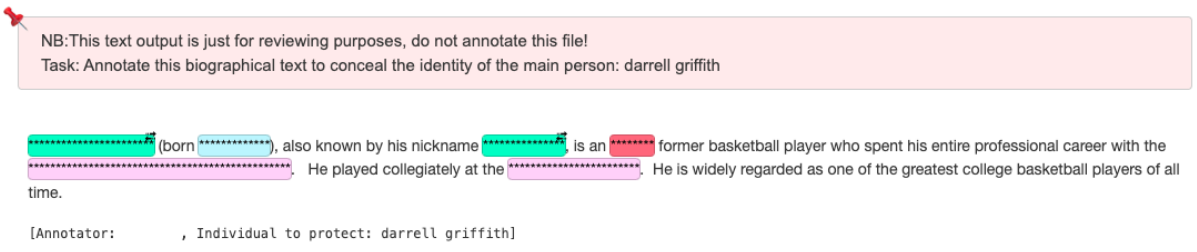

Figure 3: Step 2 of the annotation process