Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings

Fadi Hassan, David Sánchez, Jordi Soria-Comas and Josep Domingo-Ferrer
Department of Computer Science and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia
UNESCO Chair in Data Privacy
Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona,Catalonia
{*fadi.hassan, david.sanchez, jordi.soria, josep.domingo*}*@urv.cat*

***Abstract*—Data sharing is key in a wide range of activities but raises serious privacy concerns when the data contain personal information. Anonymization mechanisms provide ways to transform the data so that identities and/or sensitive data are not disclosed (i.e., data are no longer personal). Even though a variety of methods have been proposed for structured data, automatic anonymization of unstructured text it still far from being solved. Textual data anonymization consists of detecting sensitive pieces of text, which are later removed and/or generalized. The detection process is especially challenging and it is usually based on classifiers pre-trained on large quantities of manually tagged data, which are able to detect a fixed set of (sensitive) entities such as names or locations. However, this approach is severely limited because sensitive information may appear in text in many forms and not all the appearances of a certain entity type may disclose information on the individual to be protected. In this work we propose a more general solution to text anonymization based on the notion of word embedding. The idea is to represent all the entities appearing in the document as word vectors that capture their semantic relationships. Then a particular entity (e.g. an individual or an organization) can automatically be protected by removing the other entities co-occurring in the document whose vectors are similar to the particular entity's vector. Furthermore, our method does not require manually tagged training data and is language-agnostic. We empirically evaluated our proposal on a collection of biographies. Our results show a significant improvement of the detection recall in comparison with classical approaches to text anonymization based on named entity recognition.**

***Index Terms*—Document anonymization, Privacy protection, Word embeddings, Named entity recognition.**

1. Introduction

Data are a crucial resource for many businesses and researchers. For instance, in medical and pharmacological research, there is a need to access the health records of the population under study. While in some cases researchers may be able to collect the required data themselves, the amount of data collected in this way is likely to be limited. Having access to the (health) records of a wider population is paramount for many research activities to be successful. This illustrates the importance of data sharing, as it is the only way to make specific types of data available to a broad community. Of course, personal data cannot be shared carelessly. This is especially so in the case of health data, which are particularly sensitive. Steps need to be taken for the privacy of the data subjects —the people or the organizations the data refer to— to be well protected, not only to comply with data protection regulations [10], [11] but also for ethical reasons.

According to the European Union's General Data Protection Regulation (GDPR, [10]), there are two legal ways to deal with personally identifiable information (PII): either to obtain explicit informed consent from data subjects to process the PII, or to anonymize the PII to obtain data that are no longer personally identifiable and therefore not regulated by the GDPR. When collecting consent is feasible, it should be the preferred avenue, because anonymization entails utility loss. In the previous example, collecting consent from data subjects seems feasible in the context of a clinical trial, but it is rather complex when using data collected by a third party. Hence, in general anonymization is required.

In what regards data formats, the academic literature on anonymization tends to consider structured data that conform to a regular model such as a database schema. However, the vast majority of data generated nowadays are unstructured [7], [25]. Plain textual data are the most common form of unstructured data, found in accounting books, articles, web pages, emails, posts in social networks or discharge reports.

Anonymization of structured data is a relatively mature technology. In structured data bases, attributes are categorized according to the re-identification risk they entail: an identifier is an attribute whose values are enough to re-identify records, whereas quasi-identifiers are attributes that separately do not allow re-identification but whose combination may. The usual approach to data protection is to remove identifiers and mask quasi-identifiers (where masking can mean perturbing, generalizing or even remov-

ing some values). While identifier attributes are usually easy to recognize, quasi-identifiers are not. In general, we should classify as quasi-identifier any attribute whose value may be available in an external data source that associates it with an identity. Unfortunately, this condition is hard to verify: first, it is hard to account for all available data sources and, second, a source may not be available now but become available in the future.

If dealing with structured data may be challenging, anonymization of unstructured textual data is even more complex. For one thing, we no longer have a fixed list of attributes: textual data may contain any information and it varies across documents. Furthermore, deciding what is a quasi-identifier is much more complex than with structured data: for each piece of information in the text we need to judge whether it can be used for re-identification. Such judgment is not easy for a human expert [3], let alone for an computer program.

In general, accurate anonymization of textual documents remains a largely manual process. Some tools based on named entity recognition (NER) have been designed to remove some of the burden from the human expert. These tools pinpoint specific pieces of information that are likely to facilitate the re-identification of data subjects, such as names, locations or dates.

However, NER-based techniques have important limitations. First, for the more sophisticated NER techniques based on statistical models, we need to train the NER classifiers, which requires a large amount of manually tagged training data in the same language of the text to be protected. Tagging the training data may involve an enormous effort. Second, to obtain anonymized text that retains sufficient utility, the participation of a human expert remains essential. The human expert is required to assess the extent to which the pinpointed items may re-identify the subject to whom the text refers. Some of the items may not be directly related to the subject and hence could be preserved without increasing the re-identification risk: for example, in a sentence "Most people in the station took trains for Paris" the item "Paris" can probably be preserved. However, the NER system is likely to mask every detected identifier (in the example, it would mask "Paris"), which results in unnecessary utility loss. Further, while the NER system detects a fixed set of item types (like names of people or locations), the human expert is expected to find re-identifying items that the system cannot detect: for example, a sentence that contains no identifiers but gives a physical description uniquely identifying the subject.

Contributions and plan of this paper. Our main purpose is to overcome some of the limitations of NER-based techniques for textual data anonymization. To this end, we design a more general method that captures better the notion of disclosure or re-identification risk caused by the textual terms appearing in a document. Our method analyzes the semantic relatedness among textual terms by leveraging the notion of word embedding and, more specifically, its neural network-based implementation [14]. Word embedding learns vector representations of linguistic terms according to their (co-)occurrence in large corpora. We make use of word embedding to measure the relatedness and, therefore, the extent to which the terms appearing in the document to be anonymized may re-identify the entity to be protected —a personal subject or even an organization. In this way, we automatically classify the items of information contained in unstructured texts into potentially identifying or not. This configures a more powerful solution to anonymize textual documents: it is not restricted to detection of predefined types, but can limit masking only to that information that is potentially disclosive for the entity to be protected. Additionally, our solution is language-agnostic and does not require manually tagged training data. Last but not least, the empirical results we report in this paper show that our method significantly improves the detection recall on NER-based approaches and, therefore, yields more robust anonymizations.

The rest of this paper is organized as follows. In Section 2 we review related work on textual document anonymization. In Section 3 we present our approach to document anonymization based on word embedding. An empirical evaluation of our method and a comparison against related works is given in Section 4. Conclusions and future work are gathered in Section 5.

2. Related work

Document anonymization is a process that consists of the following two steps: (i) detecting pieces of information that can be used to re-identify the entity to be protected, and (ii) masking such pieces of information appropriately (that is, by applying the least amount of masking required to keep the risk of re-identification within an acceptable level). In this work, we focus on the first step, that is substantially more difficult than the second [3].

As previously said, anonymizing textual documents currently remains a highly manual process [17]. A human expert (or a group of them) is in charge of reviewing the text and masking all items that might be used to re-identify subjects. To reduce the burden of human experts, some systems that make use of named entity recognition (NER) have been introduced.

NER was created as a way to extract structured information from an unstructured text. It was noted that effective information extraction required the system to recognize entities such as person and organization names, locations, times, and dates. Early NER systems were based on handcrafted rules (or regular expressions). For instance, times can be identified using the following pattern: "at" + digits + "am"/"pm". Up until 2000, handcrafted rule systems offered the best results. Statistical approaches subsequently took over. In statistical NER systems models such as HMM (hidden Markov models) or CRF (conditional random fields) are trained to locate a specific type of entity. A supervised learning approach is used to train the models, which requires large amounts of tagged data. Well-trained NER models have high precision (typically, above 80%). Additionally, there are quite a few software packages available to carry

out automatic NER tasks, such as spaCy [26] or the Stanford NER [12].

In the context of textual anonymization, NER is used to detect information items likely to be useful to re-identify the subject. This allows performing automatic anonymization by masking —which may involve perturbing, generalizing or removing— all the detected items; alternatively, NER-based system can be used to assist a human expert that remains in charge of the anonymization. As mentioned in the introduction, automatic NER-based anonymization has a number of issues.

Simple text anonymization products using rule-based NER are recalled next. Adobe Professional [18] has a redaction feature that is capable of pinpointing the following types of entities: phone numbers, credit cards, social security numbers, email addresses and dates. While this tool is helpful, the types it detects are rather limited. The Scrub system [27] makes use of rules and dictionaries to locate and mask several types of re-identifying information items. In some domains there are specific regulations that determine how anonymization should be performed. This is the case of the safe harbor provisions established by the HIPAA (Health Insurance Portability and Accountability Act, [11]). Systems such as [9], [28] enforce the HIPAA safe harbor provisions by using dictionaries to look for common names (e.g. persons, hospitals) and rules to locate other type of data (e.g. dates).

Solutions based on statistical NER and machine learning are also available. Amazon's Macie [1] locates several personally identifiable information items (such as names, addresses, dates of birth, etc.) and classifies documents in several categories according to their risk of re-identification. Additionally, Macie is capable of detecting many information items that must be kept secret regardless of the fact that they may be linkable to a specific individual (e.g. passwords, bank accounts, etc). Google's Cloud DLP [5] makes use of rules and machine learning techniques to detect the presence of confidential and re-identifying pieces of information. Additionally, Cloud DLP can automatically perform basic anonymizations and compute risk metrics based on privacy models such as k-anonymity, l-diversity or δ-presence. Similarly, Symantec's Data Loss Prevention [8] makes use of dictionaries and rules (to detect several types of information items that have a regular structure) and machine learning (to detect other types of identifiable and confidential information that lack a regular structure).

3. The proposed method

The most widely accepted definition of privacy amounts to informational self-determination, that is, the ability of individuals, groups or organizations to seclude themselves or information about themselves selectively [29]. Following this definition, the crux of anonymizing data releases is the ability to detect (and subsequently remove or mask) the information that refers to a single entity and to no other entity. This is exactly what our approach sets out to achieve. As discussed in the introduction, approaches based on NER fail in this respect and implicitly assume that the whole content of each document exclusively refers to one single entity.

To reach our goal, we need a way to characterize the linguistic terms appearing in a document according to the information they disclose on the entity to be protected. A usual measure of this "amount of disclosure" is the degree of relatedness between the terms in the document and the entity [23]. Traditionally, the semantic relatedness between linguistic entities has been assessed using distributional [15] or probabilistic models [20], which require accurate statistics on the (co-)occurrence of words. A recent trend in computational linguistics to measure the relatedness between words is to use neural network-based word vectors or word embedding models.

Word embedding maps words into high-dimensional vectors. For the vector representation of a word to be useful, we want similar words to be mapped to similar vectors. Several ways to assign vectors to words have been proposed. The current state of the art is based on the use of 2-layer neural networks [13]. The neural network is trained either to predict the current word from a window of neighboring words (continuous bag of words) or to predict neighboring words based on the current word (skip-gram). To this end, the network uses a collection of documents or sentences as input data, and builds a vocabulary from the words appearing in the collection. The weights obtained by training the neural network for each word in the vocabulary are used as the vector associated with that word.

The skip-gram model is usually more accurate [13]. It yields output probabilities expressing how likely it is to find a word of the vocabulary in the neighborhood of the input word. From the perspective of *distributional semantics* words likely to co-occur in a context (or, otherwise put, those with similar contexts) tend to be semantically related [19]. A strong semantic relatedness between the words appearing in the text and the entity to be protected is what enables the semantic inferences that may cause disclosure [2], [23]. Due to this convenient feature, we propose to use the skip-gram word embedding model as the means of detecting the terms appearing in a document that may compromise the privacy of a certain entity.

Our approach consists of two phases. In the first phase, we use a large corpus to train a word embedding model tailored to capture the semantic relationships that may cause disclosure. This model is general and conveys the relationships (and, therefore, the pairwise disclosure risks) between all the terms appearing in the document collection. In the second phase, we use the trained model to detect the terms contained in the document to be protected that may disclose the identity of an entity. Even though building the model is costly, once built it can be efficiently reused to anonymize all the documents w.r.t. all the entities appearing in the collection.

3.1. Training the model

The first phase of our method is depicted in Figure 1. It can be described in terms of the following steps, explained further below:

- Data collection;
- Data pre-processing;
- Configuration of training parameters;
- Model training.

3.1.1. Data collection. To train a word embedding model that accurately characterizes the disclosure-enabling relationships affecting an entity we need a representative collection of documents describing the entity.

Ideally, the collection ought to contain all the documents that shall be anonymized (e.g., collections of medical records). In this way we ensure that all the words contained in such documents (and the entities to whom they refer) appear in the model's vocabulary and have associated vector representations. If this "core" corpus is small, the collection of documents can be extended with more general corpora that will provide additional evidences on the co-distribution of words and thereby mitigate the data scarcity.

3.1.2. Data pre-processing. Since semantic inferences occur at a conceptual level, rather than at a word level, we introduce a pre-processing step to create a meaningful vocabulary for the word embedding model.

Specifically, concepts and entities are referred to in a text via noun phrases rather than individual words. For example, the noun phrase "New York Times" refers to a specific entity that is completely different from the individual meaning of its words "New", "York" and "Times". Thus, to properly evaluate disclosure risks, we need the vector representation of the *concepts* referred to by the text (e.g., "New York Times"), rather than representations of the individual words. Therefore, in the pre-processing step we extract the noun phrases and feed them as training data to the word embedding system.

The pre-processing step consists of a pipeline of syntactical analyses: tokenization, part-of-speech tagging and chunking [16]. As a result, noun, verb and prepositional phrases are obtained. Also, to minimize the lexical variability of the noun phrases, stop words are removed during the tokenization step; in this way, the occurrences of phrases like "the New York Times" and "New York Times" will contribute to the same vocabulary entry/word vector.

In addition to improving the characterization of the entities mentioned in the documents, this pre-processing helps to reduce the size of the vocabulary and, therefore, the training runtime.

3.1.3. Configuring the training parameters. The standard implementation of neural network-based word embeddings is Word2Vec [14], which we use in this work. In this section we discuss the parameters involved in the configuration of Word2Vec and the training of the word embedding model in the context of document anonymization.

As said above, we use the skip-gram architecture to train the word embedding model. Skip-gram predicts the probability that words appear in the neighborhood of the input word within a fixed window size. The window size is usually set to encompass complete sentences, that is, between 5 and 10 words. Larger window sizes require more iterations (i.e., more word pairs are evaluated during the learning process); in fact, doubling the window size increases the runtime by around 50%.

Another relevant parameter is the dimension of the word vectors. In principle, the greater the dimension, the more accurate the results. However, since the dimension is equal to the number of neurons in the network, a higher dimension has a direct impact on the training runtime. Again, doubling the size of the vectors usually implies increasing the runtime by around 50%. Even though there is no fixed rule to tune the dimension, a value 300 is suggested in [13].

Finally, it is possible to set a minimum number of appearances as a cutting threshold below which the words in the training data will be discarded. Since word embedding is usually employed to guide semantic similarity assessments, it makes sense to discard words from the input corpus that occur too rarely for the model to learn accurate similarities. Also, filtering outliers significantly reduces the vocabulary size and, therefore, the training runtime. However, in the context of document anonymization, rare words (such as names or particular addresses, which would appear once) are usually those that entail a greater risk because they often refer to very specific (quasi-)identifying information [21]. For this reason, we do not use any cutting threshold for rare words.

3.2. Detecting sensitive terms

After the model is built, we obtain a vector representation of each phrase in the input collection of documents. Any two phrases with similar contexts (i.e., closely related in semantics [19]) will also have similar vectors. The standard way of measuring the similarity between vectors is the cosine similarity.

Our setting is depicted in Figure 2. Given a document to be anonymized with respect to a particular entity (e.g. denoted by its full name), we iteratively evaluate the disclosure risk of the terms in the document by measuring the cosine similarity between their vector representations and the vector of the entity to be protected. Prior to that, the document to be anonymized undergoes the pre-processing described in the first phase, so that linguistic entities are evaluated at a phrase level rather than at a word level. If the resulting similarity is above a threshold t then the evaluated term is deemed sensitive and needs to be masked.

Different strategies may be employed to mask sensitive terms, such as removal or generalization. The former strategy is straightforward and is usually employed in document redaction [22]. On the other hand, term generalization may be more desirable because it preserves better the semantics and the readability of the protected document. However, generalizing requires a detailed knowledge base from which

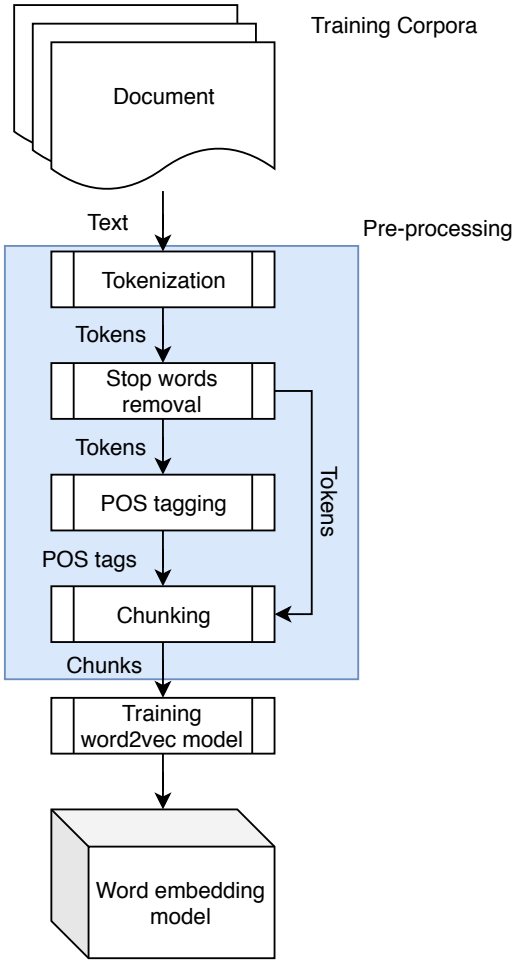


Figure 1. First phase steps

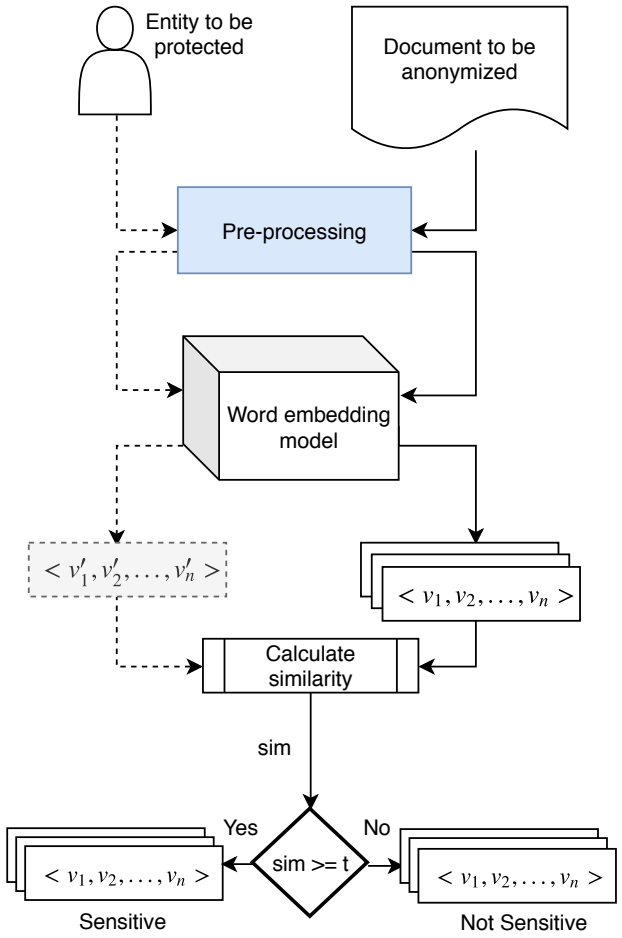


Figure 2. Second phase steps

suitable generalizations of the sensitive terms can be obtained. Moreover, determining the combination of term generalizations that optimizes the trade-off between disclosure risk protection and preservation of the document semantics is an NP-hard problem [23]. The latter has been tackled in the past by means of greedy heuristic algorithms [24].

4. Evaluation

We evaluated our method in a scenario similar to that used in related works on document redaction [4], [21], [22]. These usually employ as evaluation data a collection of Wikipedia articles describing entities of different domains. Wikipedia articles are a good choice due to their high informativeness and tight discourses, which are challenging for document anonymization. The contents of each article are then manually tagged to label the terms that may re-identify the subject described by the article.

In our case, we focused on Wikipedia English articles about movie actors from several countries. First, we collected the abstracts of 19,000 articles under the "20$^{\text{th}}$ century actors" Wikipedia category. Then, we used these data as the collection of documents to train the word embedding model as explained in Section 3.1. The contents of this collection resulted in a vocabulary of 217,019 elements. Our training parameters were the standard ones discussed in Section 3.1: window size 10 and vector dimension 300. The evaluation test bed consisted of 50 randomly picked abstracts from the collection. These were manually tagged to label words and phrases that might disclose the actor's identity. As a result of the tagging, 2,655 words were labeled as sensitive, which accounted for 30% of the content. Then, we measured the performance of our method by comparing the manual tags (which constituted the ground truth) against the terms our method detected as sensitive. Note that *we used manual tagging just for performance evaluation, not for training*.

The evaluation metrics we employed were the standard precision, recall and F1-score measures.

Precision is defined as

$$Precision = \frac{\#correct\ terms}{\#detected\ terms} \cdot 100,$$

where $\#detected\ terms$ is the number of terms detected as sensitive by the process described in Section 3.2 and

$\#correct\ terms$ is the number of terms that contain one or more manually tagged words. The higher the precision, the lower the proportion of false positives, that is, of over-masked terms. A high precision implies that the semantics and readability of the protected document are well preserved.

Recall is defined as

$$Recall = \frac{\#detected\ tagged\ terms}{\#tagged\ terms} \cdot 100,$$

where $\#tagged\ terms$ is the number of terms manually tagged as sensitive and $\#detected\ tagged\ terms$ is the number of tagged terms that are fully contained in a term detected by our method. The higher the recall, the lower the proportion of false negatives and the more robust the anonymization is.

Last, the F1 score is

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \cdot 100,$$

which corresponds to the harmonic mean of precision and recall and summarizes the performance of the method when the same weight is attached to precision and recall. Notice that, while a high precision is good to preserve the document semantics, a high recall is needed to keep the disclosure risk under control (non-detected sensitive terms may allow re-identification).

Taking the above into account, we set the similarity threshold t (see Section 3.2) so that the F1-score was maximized on average for all the documents. The selected value was $t = 0.25$. Higher threshold values resulted in a slight precision increase at the cost of a significant recall degradation. Conversely, lower threshold values did not significantly increase the recall and degraded precision.

The results of our method were compared against those of a standard NER-based approach. In particular we used the state-of-the-art Stanford Named Entity Recognizer software [12], which provides 3 pre-trained NER models for English:

- NER3: detects and categorizes named entities of ORGANIZATION, LOCATION and PERSON types;
- NER4: detects and categorizes named entities of ORGANIZATION, LOCATION, PERSON and MISC types;
- NER7: detects and categorizes named entities of LOCATION, ORGANIZATION, DATE, MONEY, PERSON, PERCENT and TIME types.

A recent study shows that the Stanford NER achieves a precision close to 90% and a F1 around 85% in pure NER tasks [6].

Table 1 reports the average results of the different methods for the 50 evaluated documents. It is clear from the figures that our method outperforms the NER-based approach very significantly, regardless of the NER model used. In particular, recall is more than doubled, which results in a much higher F1-score. This illustrates the main limitation of NER-based methods: they only detect named entities, but there are other sources of disclosure. This limitation tends to produce under-protected documents in which entities may be re-identified by correlating several (non-masked) facts or personal features that do not fall into the predefined types of named entities.

TABLE 1. AVERAGE PRECISION, RECALL AND F1-SCORE FOR THE 50 EVALUATED ARTICLES

	Precision	Recall	F1
NER3	96.09%	19.59%	32.07%
NER4	97.59%	34.25%	49.72%
NER7	**98.32%**	27.89%	42.77%
Our method	82.69%	**81.24%**	**81.66%**

TABLE 2. PRECISION, RECALL AND F1-SCORE FOR A DOCUMENT REFERRING TO TWO DIFFERENT SUBJECTS

	Precision	Recall	F1
NER 3	55.55%	22.72%	32.25%
NER 4	**77.27%**	59.09%	66.96%
NER 7	60.0%	27.27%	37.5%
Our method	68.0%	**81.81%**	**74.27%**

Precision is the only metric for which the NER-based approach achieves better results. Indeed, as mentioned above, NER is good at recognizing named entities. Moreover, the evaluation scenario we consider is particularly favorable to NER because most of the text in each document is very related to the subject to be protected. Therefore, if a named entity (which by definition is a highly specific data item) appears in the text and is properly identified by the NER method, then it is very likely to be sensitive.

In a less favorable scenario, in which the content of a document may refer to several subjects, the anonymization-related precision of the NER-based approach would significantly decrease, because not all the named entities in the document would refer to the subject to be protected. We simulated this scenario by merging the articles of two related actors (both American and acting in the same TV series) and focusing the manual anonymization on one of them. The results of this experiment are reported in Table 2. As expected, the precision of the NER-based method is now significantly lower, even though we see relevant differences among the different NER models, which we discuss below. Our method also increases its false positive rate (because of the relatedness between the two subjects), but in a smaller percentage, whereas the recall is maintained at the same level.

The behavior of the different methods is illustrated in Table 3, which contains an extract of the input text of one of the evaluated articles and compares the manual tagging with the entities detected by the different approaches. We can see that the NER-based method fails to detect pieces of information that are relevant to re-identify the actor, such as his nationality, birthdate or the title of the TV

TABLE 3. OUTPUT SAMPLES FOR EACH METHOD

manually tagged text	**Jared Tristan Padalecki** (born **July 19, 1982**) is an **American** actor. He is best known for playing the role of **Sam Winchester** in the **TV series Supernatural**.
NER3	{PERSON[Jared_Tristan_Padalecki]} (born July 19, 1982) is an American actor. He is best known for playing the role of {PERSON[Sam_Winchester]} in the TV series Supernatural.
NER4	{PERSON[Jared_Tristan_Padalecki]} (born July 19, 1982) is an {MISC[American]} actor. He is best known for playing the role of {PERSON[Sam_Winchester]} in the TV series {ORGANIZATION[Supernatural]}.
NER7	{PERSON[Jared_Tristan_Padalecki]} (born {DATE[July_19_,_1982]}) is an American actor. He is best known for playing the role of {PERSON[Sam_Winchester]} in the TV series Supernatural.
Our method	{SENSITIVE[jared_tristan_padalecki]} (born {SENSITIVE[july_19_,_1982]}) is {SENSITIVE[an_american_actor]}. He is_best_known for playing {SENSITIVE[the_role]} of {SENSITIVE[sam_winchester]} in {SENSITIVE[the_tv_series_supernatural]}.

TABLE 4. AVERAGE COEFFICIENTS OF VARIATION (CV) FOR PRECISION, RECALL AND F1-SCORE

	Precision CV	Recall CV	F1 CV
NER3	0.48%	2.14%	2.41%
NER4	0.16%	3.67%	2.97%
NER7	**0.09%**	2.65%	2.34%
Our method	0.52%	**0.67%**	**0.31%**

series (which, being an adjective, does not fall into the predefined categories of named entities). In contrast, our approach detects all the manually tagged entities, and only incurs over-masking for the term "the role". The differences among the three NER models are also quite illustrative. Neither NER3 nor NER4 are able to detect dates (such as birthdates, which are highly disclosive), because they have not been trained to detect this type of alphanumerical entities. Also, whereas NER3 and NER7 have been trained with specific named entity types, NER4 also detects the generic MISC type, which includes named entities such as nationalities. As shown in Tables 1 and 2, thanks to MISC the NER4 model obtains the best recall among the NER-based models, followed by NER7 (that detects the greatest number of specific entity types) and, finally, NER3.

For a method to be useful it has to yield not only good average results, but good enough results in all documents. Table 4 reports the coefficient of variation (a measure of dispersion computed as the ratio between the standard deviation and the mean) of the results in Table 1. We can see that our approach provides the most stable results, with a variation of the F1-score of just 0.31%.

5. Conclusions and future work

Document anonymization has been traditionally conducted (or at least assisted) by trained classifiers that detect predefined types of entities. Such methods assume that all entities of a certain type are sensitive whereas the remaining text is not. This is a very coarse approach that typically results in under-protection for terms that do not fall into the predefined types, as we have empirically shown in the previous section. Restricting anonymization to predefined types lacks flexibility to detect text descriptions that might be re-identifying. Finally, in order to detect named entities, manually tagged training data are needed that match the language of the text to be protected, and this implies a significant human effort.

In contrast, our method is more general and, at the same time, more flexible. First, it is not restricted to pieces of text with "regular" structures (named entity types) because *all* the terms and phrases appearing in the text are evaluated, and the disclosure they cause is assessed according to the anonymization requirements (i.e., preventing re-identification of the subject to be protected). This approach is more similar to the manual sanitization process performed by human experts. Thus, it is natural for our method to significantly improve the detection recall, thereby yielding more robust anonymization. Finally, even though we rely on a word embedding model built from training data, such data are not manually tagged and the construction of the model is language-agnostic. Therefore, no manual effort is required during the whole lifecycle of the protection process, which makes our method suitable for managing large amounts of textual data.

As future work, we envision several ways to improve our results. First, we can improve the pre-processing step by incorporating morphological analyses by which derivative forms of the same words/phrases (e.g., singular/plural) can be identified and aggregated; this will contribute to reducing the size of the vocabulary and hence improving the training efficiency; further, it will capture more accurately the distribution of the entities referred to by the text. Second, we plan to design sanitization algorithms that generalize —rather than remove— sensitive entities consistently with the detection threshold; to this end, we plan to use large knowledge bases (e.g., YAGO or WordNet), whose concepts will also be incorporated into the word embedding model. Additional experiments will also be considered that leverage not only domain-specific documents, but also general-purpose corpora (e.g., DBPedia or the Google News data set).

Acknowledgments

This work was partly supported by the European Commission (project H2020-700540 "CANVAS"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705) and the Spanish Government (projects TIN2014-57364-C2-1-R "SmartGlacis", TIN2015-70054-REDC, RTI2018-095094-B-C21

and TIN2016-80250-R "Sec-MCloud"). While the authors are with the UNESCO Chair in Data Privacy, the opinions expressed in this paper are the authors' own and do not necessarily reflect the views of UNESCO.

References

[1] *Amazon Macie*, Amazon web services (aws). Accessed May 16, 2019. https://aws.amazon.com/macie/

[2] B. Anandan and C. Clifton, "Significance of term relationships on anonymization," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 3, pp. 253–256, IEEE Computer Society, 2011.

[3] E. Bier, R. Chow, P. Golle, T. H. King, and J. Staddon, "The rules of redaction: identify, protect, review (and repeat)," *IEEE Security & Privacy*, vol. 7, no. 6, pp. 46–53, 2009.

[4] R. Chow, P. Golle, and J. Staddon, "Detecting privacy leaks using corpus-based association rules," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 893–901, ACM, 2008.

[5] *Cloud data loss prevention*, Google Cloud. Accessed May 16, 2019. https://cloud.google.com/dlp/

[6] C. M. Costa, G. Veiga, A. Sousa, and S. Nunes, "Evaluation of Stanford NER for extraction of assembly information from instruction manuals," in *2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pp. 302–309, IEEE, 2017.

[7] "Data, data everywhere: A special report on managing information," *The Economist*, Feb. 27, 2010.

[8] *Data Loss Prevention,* Symantec. Accessed May 16, 2019. https://www.symantec.com/products/dlp

[9] M. Douglass, G. Cliffford, A. Reisner, W. Long, G. Moody, and R. Mark, "De-identification algorithm for free-text nursing notes," in *Computers in Cardiology 2005*, pp. 331–334, IEEE, 2005.

[10] *General Data Protection Regulation*. European Union Regulation 2016/679, 2016. https://gdpr-info.eu

[11] *Health Insurance Portability and Accountability Act*, 104th U.S. Congress Public Law 191, 1996. https://www.govinfo.gov/content/pkg/PLAW-104publ191/html/PLAW-104publ191.htm

[12] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

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

[15] S. M. Mohammad and G. Hirst, "Distributional measures of semantic distance: A survey," *arXiv preprint arXiv:1203.1858*, 2012.

[16] T. Morton, J. Kottmann, J. Baldridge, and G. Bierner, "OpenNLP: A Java-based NLP toolkit," in *Proc. of 10th Conf. of the European Chapter of the Association for Computational Linguistics (EACL 2005)*, 2005.

[17] National Security Agency, *Redacting with confidence: how to safely publish sanitized reports converted from Word to PDF*, Report #I333-015R-2005, Dec. 13, 2005.

[18] National Security Agency, *Redaction of PDF files using Adobe Acrobat Professional X,* Report #I73-025R-2011, Nov. 22, 2011 (reviewed Jul. 16, 2015).

[19] M. Sahlgren, "The distributional hypothesis," *Italian Journal of Disability Studies*, vol. 20, pp. 33–53, 2008.

[20] D. Sánchez, M. Batet, A. Valls, and K. Gibert, "Ontology-driven web-based semantic similarity," *Journal of Intelligent Information Systems*, vol. 35, no. 3, pp. 383–413, 2010.

[21] D. Sánchez, M. Batet, and A. Viejo, "Automatic general-purpose sanitization of textual documents," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 853–862, 2013.

[22] D. Sánchez, M. Batet, and A. Viejo, "Minimizing the disclosure risk of semantic correlations in document sanitization," *Information Sciences*, vol. 249, pp. 110–123, 2013.

[23] D. Sánchez and M. Batet, "C-sanitized: A privacy model for document redaction and sanitization," *Journal of the Association for Information Science and Technology*, vol. 67, no. 1, pp. 148–163, 2016.

[24] D. Sánchez and M. Batet, "Toward sensitive document release with privacy guarantees," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 23–34, 2017.

[25] C. C. Shilakes and J. Tylman, *Enterprise Information Portals*, Merrill Lynch Inc., New York, 1998.

[26] *spaCy: Industrial-Strength Natural Language Processing in Python.* Accessed May 16, 2019. https://spacy.io

[27] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system.," in *Proceedings of the AMIA Annual Fall Symposium*, p. 333, American Medical Informatics Association, 1996.

[28] A. Tveit, O. Edsberg, T. Rost, A. Faxvaag, O. Nytro, T. Nordgard, M. T. Ranang, and A. Grimsmo, "Anonymization of general practioner medical records," in *2nd HelsIT Conference (HelsIT 2004)*, 2004.

[29] A. F. Westin, *Privacy and Freedom*, Atheneum, New York, 1967.