Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence *

Fabio Martinelli*, Fiammetta Marulli†, Francesco Mercaldo‡*, Stefano Marrone†, Antonella Santone§
*Institute for Informatics and Telematics, National Research Council of Italy (CNR), Pisa, Italy
{fabio.martinelli, francesco.mercaldo}@iit.cnr.it
†Department of Maths and Physics, University of Campania "L. Vanvitelli", Caserta, Italy
{fiammetta.marulli, stefano.marrone}@unicampania.it
‡Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, Campobasso, Italy
francesco.mercaldo@unimol.it
§Department of Biosciences and Territory, University of Molise, Pesche (IS), Italy
antonella.santone@unimol.it

***Abstract*—Artificial Intelligence systems have enabled significant benefits for users and society, but whilst the data for their feeding are always increasing, a side to privacy and security leaks is offered. The severe vulnerabilities to the right to privacy obliged governments to enact specific regulations to ensure privacy preservation in any kind of transaction involving sensitive information. In the case of digital and/or physical documents comprising sensitive information, the right to privacy can be preserved by data obfuscation procedures. The capability of recognizing sensitive information for obfuscation is typically entrusted to the experience of human experts, who are overwhelmed by the ever increasing amount of documents to process. Artificial intelligence could proficiently mitigate the effort of the human officers and speed up processes. Anyway, until enough knowledge won't be available in a machine readable format, automatic and effectively working systems can't be developed. In this work we propose a methodology for transferring and leveraging general knowledge across specific-domain tasks. We built, from scratch, specific-domain knowledge data sets, for training artificial intelligence models supporting human experts in privacy preserving tasks. We exploited a mixture of natural language processing techniques applied to unlabeled domain-specific documents corpora for automatically obtain labeled documents, where sensitive information are recognized and tagged. We performed preliminary tests just over 10.000 documents from the healthcare and justice domains. Human experts supported us during the validation. Results we obtained, estimated in terms of precision, recall and F1-score metrics across these two domains, were promising and encouraged us to further investigations.**

***Index Terms*—Privacy, Data Protection, Natural Language Processing, Sensitive Data Extraction, Artificial Intelligence, Unsupervised Machine Learning**

I. Introduction

It is commonly recognized that Privacy and Data Protection (PDP) offers a cutting edge in legal, regulatory, academic and technological development [1], [2]. It is also well known that Artificial Intelligence (A.I.) has offered new methods and solutions in various application fields [3], thus enabling significant benefits for users and society.

Artificial Intelligence and machine learning (M.L.) systems work better as larger are the data sets by which they can be fed, trained and tuned. Big data represent a relevant source of data which are currently provided by the most diverse and heterogeneous sources but, whilst the data increase in quantity, they decrease in their quality. Furthermore, Big data and, more generally all the data that can be spread and circulate among users without any control, offer the side to privacy and security leaks. In fact, such data contain more and more sensitive and personal information, without any possibility for the respective owners to regulate the access and grant the required permissions for accessing to it.

The rapid development of the information and communication technologies delivered unheard of quantities of information to people, but introduced severe vulnerabilities to the right to privacy, to the point that governments were obliged to enact more specific regulations to ensure privacy preservation in any kind of transaction involving sensitive information. An example is provided by the European Union (EU) General Data Protection Regulation (GDPR) [4], introduced in 2018 and applied across all the Members of EU.

However, even if privacy and data protection should be somehow protected by such kind of regulations, practices and frameworks for guiding the effective implementation of these regulations are still missing [5].

In the case of digital and/or physical records and documents comprising sensitive information, the right to privacy can be preserved by applying some obfuscation procedures, as the anonymization and pseudo-anonymization procedures.

Furthermore, the capability of recognizing and distinguishing sensitive information required for obfuscation tasks, is almost fully entrusted to the experience of human domain experts, who are overwhelmed by the enormous and always increasing amount of documents that require to be processed.

In the case of the privacy protection, correct and exhaustive identification of sensitive data is a very hard task, because these information range in the space comprising all the direct and indirect informational connections that can lead unauthorized subjects to trace back the identity of a person. In addition to personal data, also information relating to sexual and religious orientations, to family, for example, are considered

*Corresponding Authors: F. Marulli and F. Mercaldo

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

as private and need to be preserved.

Furthermore, each application field, can introduce some additional and domain-specific sensitive information. In this perspective, the picture gets more complicated by the fact that all the general and domain-specific knowledge for managing in the appropriate way persons' sensitive data is an heritage that only human experts possess.

Automatic and semi-automatic systems, evenly based on artificial intelligence and advanced machine learning techniques, could be proficiently employed for relieving the great effort required by these kinds of activities. These systems would be able to speed up the processing of documents and lighten the load of human officers.

Anyway, as long as this knowledge will be an exclusively heritage of humans experts and it won't be transmitted and represented in a machine readable way, artificial intelligence and machine learning won't be able to support, with the adequate effectiveness and success, such a kind of processes.

Currently, the lack of such knowledge representation and transferring represents one of the main weakness point in the development and employment of A.I. and advanced machine learning systems for supporting data and privacy protection activities.

Taking the view from the side of the current challenges triggered by the development of artificial intelligence and machine learning systems, there is the need to design and adopt international approaches and standards, in order to ensure the promotion and protection of human rights in all digital developments at international level.

So, the study described in this work is part of this problem. It proposes a sort of transfer knowledge methodology aiming to leverage knowledge across general and specific-domain tasks. Our goal is the building of knowledge bases for feeding artificial intelligence classification systems. In particular, we leverage a mixture of natural language processing techniques, based on artificial intelligence models trained over general knowledge domains and, unsupervised machine learning techniques, applied to specific domain resources. In this way we were able to produce, starting from a corpora of unlabeled documents, the corresponding labeled corpora, in which sensitive information have been recognized and tagged. This can offer a support to human experts in their privacy preserving task. Finally, we performed a preliminary test campaign including just over 10.000 documents, picked up from the healthcare and justice domains. We asked to professionals, as lawyers and doctors, to help us in the validation process of the results of our procedure. We adopted a corpora comprising 1000 documents (divided as half from the healthcare and half from the justice fields) and we performed 10 repetitions of the same procedure, by varying some clustering parameters and we computed precision, recall and F1-score, for each of the considered case study. Preliminary results were interesting and offer several hints of investigations, in order to make this methodology more accurate, robust and applicable to different domains. We were so encouraged to continue experiments over more documents and more different domains, and studying the biasing of this methodology from the tuning of some parameters, as it will be explained in the result section.

The rest the paper is organized as follows: Section II describes the most relevant related works; Section III introduces our methodology and section IV details the process implementing this methodology. The Natural language processing and the underlying complementary artificial intelligence and unsupervised machine learning techniques adopted are explained in more details. Section V describes the tools we adopted for performing preliminary tests on two case studies, represented by the healthcare and justice domain fields. Section VI concludes the work by summarizing the results obtained by the preliminary test campaign and discussing the future direction for improving this research.

II. Related Works

Among the main difficulties in implementing effective privacy and data protection countermeasures, there are the complexity in organizing comprehensive data protection plans, able to cover the mandatory prescriptions from governments regulations and the limited resources available from entities and organizations to actuate with effectiveness a data sanitization campaign. The criticality of sensitive and privacy information protection in all the transactions among systems and organizations was fully recognized by all the world governments that enacted, in the last few years, new policies for regulating privacy management, considering that large amount of data are produced and need to be revised, like the General Data Protection Regulation (GDPR), applied since 2018 in a mandatory way in all the member countries of the European Union [1]. As suggested in [4], a Data Protection Management System might be the optimal solution for ensuring consistent data protection; anyway, it was not always feasible for entities and companies, because of limited entity's budgets and resources. Thus, entities and organizations are arranging their limited resources for addressing with priority most critical issues and high-risk processing activities. Some mitigation strategies concerning data privacy preservation activities [6] in digital materials, are represented by data encryption, data anonymization [7] [8] and data pseudo-anonymization, especially applied, but not limited, to critical domains like healthcare [9] and justice. Data encryption, in the family of as homomorphic encryption techniques [10] would be suitable for managed obscured data as they were in plain but the effective application of these techniques still requires too much computational resources, thus resulting not yet applicable, not only for limited resources organization.

Unfortunately, even if the application of these mitigation actions will be able to produce significant improvements towards the data and privacy protection prescriptions, they result in being yet too much expensive for being applied with effectiveness.

Data encryption, anyway, is a particular kind of pseudo-anonymization process, since it would be possible to return

[1] https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-does-general-data-protection-regulation-gdpr-government

back to the original data, differently from the anonymization in which the correspondences between original and substitute elements is lost, with no possibility to turn back. In [11] security and privacy in Big Data cyber-physical systems is discussed, thus analyzing k-anonymization techniques, the perturbation strategies and finally homomorphic encryption applied to Big Data on healthcare domain.

Anyway, most of Public and Government (as for the case of Italy) are not ready yet to answer in an effective way to this important challenge, since they have no availability of high performance computing systems able to support cryptography processing of big amounts of sensitive data that daily they have to manage. In such a perspective, sensitive data discovery and processing in textual documents could be successfully supported by advanced Natural Language Processing (NLP) systems; in [12] discusses some mitigation strategies, based on with Laplace noise, to the problem of privacy in data centers, whose criticality is due to the device heterogeneity, by applying a local differential privacy-based classification algorithm for data centers.

In the case of Artificial Intelligence and machine learning systems for supporting privacy protection very few are the effective contributions. Whilst current literature is mainly focused on improving data anonymization or pseudo-anonimyzation techniques and research discussions are blocked on how to improve the results of data obfuscation, both in terms of techniques performances and costs and in terms of storaging of data in distributed and vulnerable to external attacks environments (e.g., cloud computing and IoTs environments), o discussions are focused on the underlying problem for arriving to obfuscation processes:

- 1. the lack of standard frameworks allowing to detect, as first, all the sensible information, in an independent and in a domain-dependent way
- 2. the lack of a common way for representing and transferring the human experts knowledge to machines, in order to support preliminary steps for detection of sensitive information.

Indeed, AI based systems, as a sensitive information classifier, need to be trained with a great amount of samples, in order to reach acceptable performances (between 80% and 90%). For generating these samples, a great effort by domain human experts is required, in the first phase of hand-crafted features in textual documents (hand-made annotation of textual documents) and in the final phase of validation of the results automatically obtained by trained systems. The scarcity of labeled corpora and the great effort required to domain experts, represents, as for each of AI and machine learning process, a great weakness. This issue is faced in this work, where we suggest a proposal for a methodology able to extract specific-domain knowledge by the scratch, exploiting the same resources that require to be processed, in order to obtain a domain knowledge in a machine readable format that could be used for training artificial intelligence and advanced machine learning systems.

III. The Proposed Methodolgy

We suggest a novel methodology for supporting human experts in the activities preceding the data obfuscation, which consists in providing an automatic process for detecting and suggesting the presence of sensitive information comprised in textual documents. Th proposed methodology is mainly based on Natural Language Processing (NLP), unsupervised and transfer learning techniques [13] and it works under the hypotesis that no external domain-specific knowledge, evenly in the shape of supplementary resources (like, for example, domain vocabularies, thesauri and/or domain ontologies, is available. To be clearer, no labeled or tagged data set and resources are available for training any any learning model for recognizing sensitive information in a text document. The general schema for the process describing our methodology is provided in Figure 1. We suppose that human expertise will be brought in the loop, only in the final stage, for validating the results obtained with the process we designed for implementing our methodology. The only resource that we suppose to have is represented by the large amount of unlabeled and plain-text documents that need to be analyzed for further obfuscation treatments. Transfer learning is used to transfer a general domain knowledge to a specific domain. In particular, we exploit general knowledge about the possibility to recognize in a textual documents some specific identifiers, as the proper names of persons, places, organizations and some other identifiers that can be considered as transversal to all the domains. This general knowledge, regardless the specific application domain, can be transferred for extracting more specific knowledge from domain documents. Finally, unsupervised learning is adopted for extracting and fine tuning domain-specific knowledge, by mining data offered by the domain corpora.

In such a perspective, the proposed methodology exploits general linguistic features for extracting some elements of interest. In particular, by the means of typical NLP pipeline, we are interested to extract Named Entities (NEs) and to perform a refining procedure for establishing if a Named Entity is candidate to be a sensitive information. This refining operation, that leads to label NEs with a more specific identifier, is performed by analyzing the context words window, surrounding the NE under examination.

In other words, we try to better characterize the semantic role of a Named Entity by analyzing its fellows words. In order to perform this contextual analysis and establishing some semantic characterization for the NEs, we need of some domain-specific knowledge. Since our working hypotesis is that we are no provided of external well assessed hand-crafted resources, we just try to extract and build this domain-specific knowledge by adopting, once again, typical processes for knowledge processing and transferring from the NLP application field. We exploit the distributional models of words within a large documents corpora, obtained by exploiting the well known Word Embeddings algorithms [14] [15] and by combining the words distribution information with topics

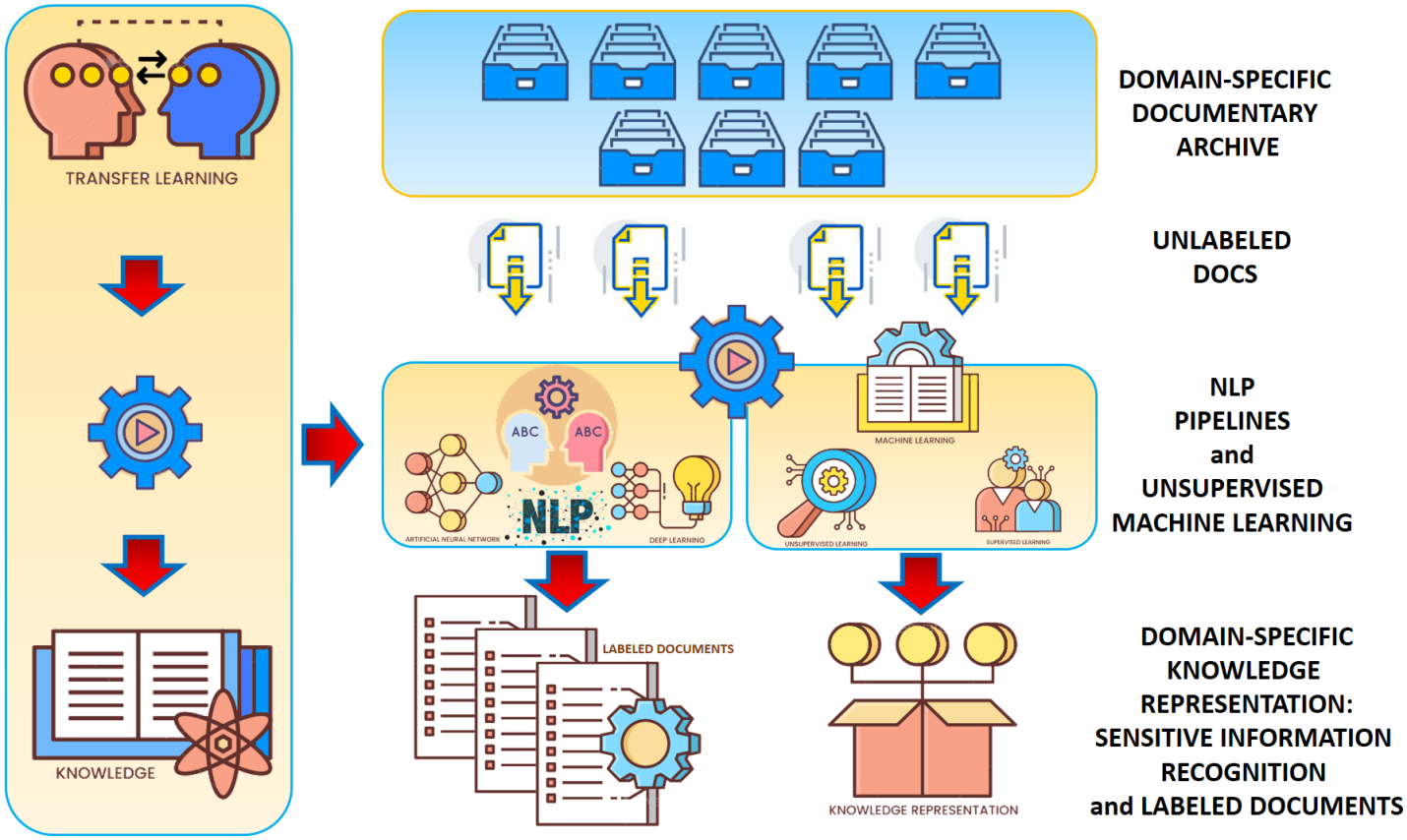

Fig. 1. An overview of the Process for Domain Knowledge Extraction and Transferring

extracted by performing a topic extraction procedure over all the available corpora of unlabeled documents. In such a way, we try to maximize the knowledge hidden in the unlabeled domain-specific documents and we try to transfer the only knowledge we have in a general context to a more specific context. The methodology we propose can be summarized as consisting of a process model whose aim is the one to produce a labeled data set for a specific knowledge domain that could be used for training a machine learning/deep learning model for automatically classifying sentences in textual documents containing sensitive information. The working hypotesis is that such a labeled data set is built from scratch, and the human expertise is required only at the end of the extraction knowledge loop, in order to validate the results that the automatic process has been able to produce by itself. Finally, this methodology was supported by the design of a process whose details are provided in the following sections.

IV. Process Details and Components

The process that implements our methodology consists in a compound pipeline of processing steps transforming each single unlabeled plain text document provided as input in the corresponding labeled version, provided as output. This process can be divided into two main phases:

- Phase 1. general knowledge transferring and domain specific knowledge extraction (*Knowledge Extraction*)
- Phase 2. domain-specific Knowledge refining and Semantic Annotation (*Knowledge Fine Tuning*)

The annotations automatically generated by the pipeline are finally submitted to a validating process, performed by human experts of the interest domain, that will provide validated versions of the labeled documents. After validation, each checked document will be included into a new data set, later adopted for training a deep learning based classification system [16] for detecting sensible information in further novel plain text documents.

A. Knowledge Extraction

Knowledge extraction is performed mainly by executing three Natural Language Processing pipelines working in parallel, allowing the domain-specific knowledge extraction and the extraction of the Named Entities that will be submitted to the final classification process for assigning appropriate labels as belonging to sensitive categories. These pipelines as showed in Figure 2. As first, from the unlabeled domain-specific corpora all the Named Entities and the topics are extracted for each single document and singularly stored for further refining analysis.

From the whole corpora of available documents, after a text normalization pre-processing step (consisting in deleting stop words and performing lemmatization of verbs), a word

embedding model is extracted by adopting a skip-gram [15] [**?**] algorithm for predicting context words.

The steps composing each of this pipelines will be briefly described in the following.

1) PoS-Tagging and Dependency Parsing: The first pipeline aims to extract Named Entities from each processed document. In order to perform this task, this pipeline comprises a Part of Speech Tagging (PoS-Tagging), a Dependency Parsing (DP) and Named Entity Recognition (NER) tasks. Several strategies and tools are available for performing both the PoS-Tagging and the Dependency Parsing [17], in several languages beyond the English. In the case of the Italian, from the recent literature are available several customized systems for accomplishing specific features of the Italian language, as the ones provided in [18] for PoS-Tagging and in [19] [20] both for PoS-Tagging and Dependency Parsing. The Named Entities Recognition task is described in more details, because it represents the first step effectively addressing the goal of the whole process.

2) Named Entities Recognition: Named entities represent identifier characterizing the identity of an entity (e.g. the name and surname of a person, the name of an organization), places, localities and unique identifiers or codes. The NEs are among the pivotal elements of a sensitive information detection process, since they are representative of those information that could incur in the need of being anonimyzed. Healthcare and justice domains provide relevant examples of field in which documents containing sensible information have to be protected from unauthorized accesses, since these documents are managed by several different agents. As first, proper names and persons' generalities (also referring to religion and sexual orientation, e.g.) need to be obscured in order to avoid abuse against individuals' privacy. The NER provide a useful mean by which a list of known named entities can be easily recognised in a text snippet, evenly supported, in this recognition process by well-known named entities lists provided to a NER system and by information provided by a Pos-Tagging and Dependency Parsing activities. Most of the NER systems provided from NLP literature are based on machine or deep learning or stochastic and probability models. All of these systems offer an annotation layer, composed of a set of predefined labels or tags that can be used for annoting named entities. In the case of the Italian language, the most performing systems are represented by Tint [19] and spaCy [20]. The set of labels provided by NER tools can be extended with further information of interest for the specific analysis and domain.

3) NER Context Window: Let we consider an textual snippet example, in the Italian language, from a clinical report about a female patient, and the NER performed by using TINT system, as in Figure 3 :

We need to distinguish better the 3 NEs detected, tagged with the "PER" label, as a patient, a doctor and a patient's related. We can analyze the "context words window" of each NE. So, for each named entity identified, we extract a context windows ($w - words$, a pre-established but variable number w of words preceding and succeeding the named entity under analysis. We we look for more information about the named entity by looking for its neighbours. In such a way, we could filter in an automatic way what to put in the candidate list for obfuscation and what to keep unaltered in the text.

Anyway, in order to look for information to filter the right named entities, we need to have a set of "sensitive words" able to evidence the role of a named entity in a text snippet. In other words, we need to be provided with a list of terms suggesting the presence of sensitive terms, that is to say we have go back to the original problem, since we lack of additional domain-specific resources.

To fill in this gap, we leveraged semantic distribution models of words, namely word embeddings models, combined with topics extracted from documents corpora. Both the word embeddings and the topic extraction, are performed as unsupervised learning processes. We built word clusters of semantic categories by adopting topics extracted from the whole corpora as categories and by populating these clusters by inferring the word embeddings model, for retrieving the neighbours words. These group of words will represent the cluster of words that we will compare with the precise context window of a named entity. Finally, we analyzed the context window foe each NE and we assigned to it the category according to the best matching among the neighbour words of the NE under analysis and the words comprised in each cluster.

In such a way, in a totally unsupervised way, we could solve the original ambiguity with no other given resources beyond the textual document to submit for data anonymization.

4) Transfer Learning and Word Embeddings Computing: Our aim is to exploit as much as possible the implicit knowledge included in the documents corpora, thus minimizing the involvement of human experts in this stage. In order to do this, we extracted knowledge included the documents corpora by training a word embeddings model [**?**] [15], able to provide us a context-specific model carrying out information about words distributions, word frequencies and words pattern co-occurrences. We exploited these context-specific word embeddings for expanding a domain vocabulary, when provided by experts and for building a new one starting from scratch, when no resources are available.

5) Topics Extraction: Topic modeling is one of the more complicated methods to identify natural topics in the text. A prime advantage of topic modeling is that it is an unsupervised technique. so, labeled training data set are not required. One of the most popular methods for topic extraction is represented by the latent Dirichlet allocation (LDA). The premise of LDA is that each text document comprises of several topics and each topic comprises of several words. The input required by LDA is merely the text documents and the expected number of topics [2].

B. Knowledge Fine Tuning

In the second phase, the rough domain-specific knowledge has to be fine tuned in order to provide some characterizing

[2]urlhttps://blog.aureusanalytics.com/blog/5-natural-language-processing-techniques-for-extracting-information

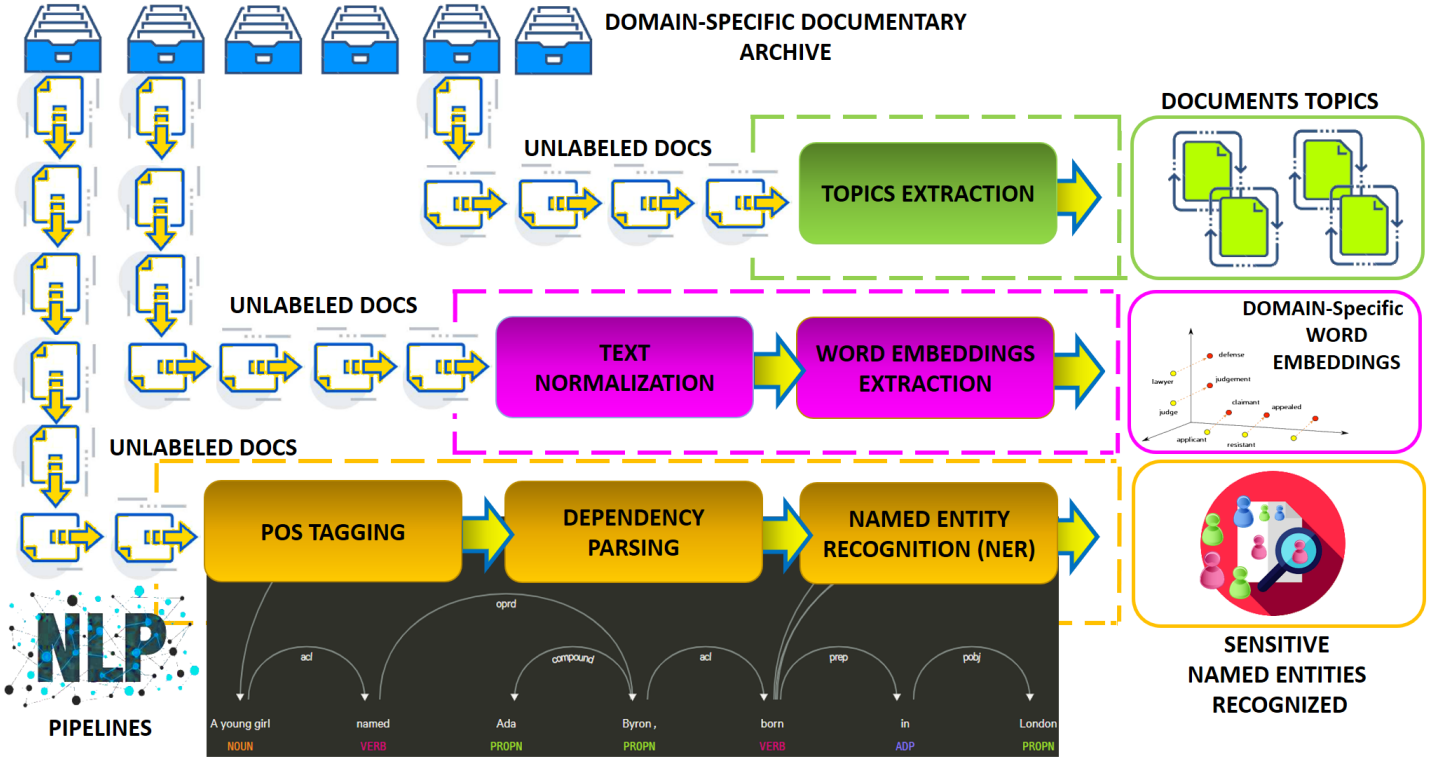

Fig. 2. The NLP Pipeline for Knowledge Extraction

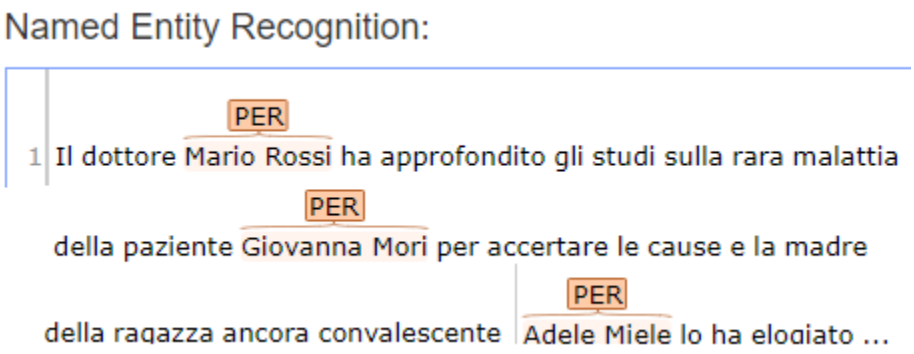

Fig. 3. Named Entity Recognition Example

features, as the sensitive categories, for distinguishing domain specific and sensitive words from all the rest and providing the final means for labeling Named Entities. The most relevant goal of this step is to extract semantic categories for sensitive information, briefly called "sensitive semantic categories". This goal is performed by combining the semantic distributional information about words, as their frequencies and co-occurrences in the whole corpora of provided documents and the topics extracted from the whole corpora (also characterized by their inverse term frequency and the relevance indicators overall the whole corpora).

By combining these knowledge, we were able to extract semantic categories. In particular, we decided to extract the n most relevant topics provided by the topic extraction process, and to build n clusters or classes. After extracting this n classes, we populated each class by the first k neighbours words to the nth topic word, extracting them from querying the word embedding model trained over the same corpora.

Finally, standing the NEs obtained from the first stage, our methodology proposes to refine NEs annotation by exploiting the extracted domain knowledge, in order to better characterize the NEs and suggesting more information to human experts during the final decision for proceeding in sensitive data management.

V. Case Studies and Tools

Finally we provide brief details for preliminary tests performed by adopting the prototyping tool implementing the

proposed methodology. We considered two case study, from healthcare and justice domains. Documents for performing our experiments were extracted, for the justice domain, from a public archive [3] provided on the web by the Italian Court of Cassation (Corte di Cassazione) and related to the 2019. For the healthcare, examples of clinical folders were provided by a professional in general medicine and were related to the past 5 years, into the district of the region Campania, in the Southern of Italy (we signed a confidentiality agreement for privacy protection of patients (that where pseudo-anonymized by our provider)).

A. Textual Corpora Description

Both the healthcare and justice corpora comprised 5000 documents, for a total amount of 10.000 documents. The language of all the provided documents was the Italian. From a lexical perspectives, each document from justice was on average composed by about 2500 words, with the shortest document counting 2213 tokens and the longest one counting 2875 tokens. Documents from healthcare domain were shorter than the corresponding from justice; they were on average composed by about 1500 words, with the shortest document counting 1179 tokens and the longest one counting 1906 tokens. We repeated twice the same steps for each of the different domain.

B. Experiment Design and Tools

The preliminary experimentation we performed was designed as follows:

1) we trained two context-specific word embeddings models over the whole corpora of documents (5000 for justice and + 5000 for healthcare)
2) we randomly selected 10% documents for testing (500 for each domain)
3) we performed by hand, with the support of domain experts, the validation of the 1000 (as 500 for justice and 500 for healthcare) labeled documents with our automatic process.

For training a context-specific word embeddings model, we adopted the Facebook *fasttext* [15] tool. In particular, we adopted a skip-gram model for learning words representation and distribution over the whole documents corpora, with embedding vectors dimension set to 150 (we adopted the skip-gram model provides better accuracy for rare and infrequent words). For performing topic extraction, we adopted a free demo version of the Text-Razor API [4]. This tool is available for 12 different languages and offers the possibility to be customized; anyway, we used the settings provided by the free demo tool, since in the demo version no customization were allowed.

For executing the natural language processing (NLP) pipeline, we adopted the spaCy tool for executing the PoS-Tagging, the Dependency Parsing and the Named Entity Recognition tasks. Instead of using the pre-trained models provided with spaCy, we trained a new model for the Italian, adopting the Italian data set provided by the Universal Dependencies Project [5] (U.D.) v2.1.

After completing the word embeddings model training, the topic extraction and the NLP pipeline over the selected 10000 documents (5000 for each domain), we proceeded with the knowledge fine tuning, by querying the word embedding trained models: we selected the size for the number of categories n = 30 and the value for the size k = 50, for each cluster. We adopted the same values for both the domains.

In both the application fields, on average, we observed that too small values of this parameters affects negatively the further semantic annotation process; a remarkable improvement is obtained when the number of clusters n is at least equal to 15 and we observed some improvements into the semantic accuracy, because we were able to better characterize the kind of information under analysis. For the until reaching a peak value to about 30. For the size of the cluster, we observed that exceeding the peak of 50, a remarkable increase in the false positives (wrong information labeled as sensitive) is recorded. The values we adopted for sizing both the number of relevant categories and the number of neighbours was obtained after repeating 10 times (10 repetitions) the experiments, for each domain.

C. Evaluation Metrics and Results Discussion

After performing the validation process, we were interested to compute precision recall and F1-score metrics over the analyzed pool of 1000 documents. Results estimated per 500 documents from the justice domain and the healthcare, are listed in table I. (*) symbol means that the 30% was erased because consisting of meaningless stop words.

TABLE I
RESULTS OF VALIDATION PROCESS FOR JUSTICE (J) AND HEALTHCARE (HC) DOMAIN

Evaluation Metrics	**Value (J)**	**Value (HC)**
Total Words Count in each document	2500	1500
effective meaningful words*	1750	1050
sensitive items included in each document	60	36
"sensitive" items detected	48	24
TRUE POSITIVES	38	18
FALSE POSITIVES	10	6
FALSE NEGATIVES	22	18
Precision (P) $p = TP/(TP+FP)$	0.79	0.75
Recall(R) $r = TP/(TP+FN)$	0.63	0.53
F1-Score $f1 = 2*P*R/(P+R)$	0.70	0.62

Since the number of sensitive categories, we grouped the results of validation process considering only two categories, the "sensitive" and the "insensitive" ones. All the sensitive fine-grained categories were grouped into a unique gross-grained category (the "sensitive" category). According to this binary differentiation, we computed two quantitative metrics, the precision, the recall nad the F1-score for estimating the

[3]http://www.cortedicassazione.it/corte-di-cassazione/it/recentissime_corte.page

[4]https://www.textrazor.com/

[5]https://universaldependencies.org/

sensitiveness and the accuracy exhibited, on average, by the proposed methodology, over the set of 500 documents automatically annoted and handly validated.

VI. Conclusions and Future Works

This work propose a methodology aiming to leverage knowledge transferring across general and specific domains for supporting the building of labeled data set useful for training A.I. and advanced supervised machine learning algorithms for sensitive privacy information detection and classification, in textual documents. Our aim is to provide an effectively working system for relieving efforts of human experts, from several application domains, during the application of data and privacy protection procedures, based on detecting and appropriate management of sensitive and personal information.

The proposed methodology is based on a mixture of natural language processing and unsupervised machine learning techniques, as topic extraction and distributional semantic models, namely word embeddings, for filling in the lack of external additional specific-domain resources, evenly for training supervised machine learning systems and providing automatic procedures for supporting human officers and experts.

We considered two case study, the healthcare and the justice in the case of the Italian language. We performed a preliminary test campaign including just over 10.000 documents, comprising of 5000 documents from the healthcare and justice domains, respectively. We performed tests over 1000 textual documents, 500 documents for each application domain and we performed a validation stage with the help of domain experts. We evaluated precision recall and F1-score for each test set, obtaining promising results.

The proposed methodology was tested for the Italian language but it can be extended to several others languages, since it only requires to change the underlying linguistic knowledge, in order to make possible transferring knowledge process across general and specific domain. Furthermore, this methodology has been tested on a reduced set of documents (1000 documents, as the total) because of the effort required in the validation stage, for fixing hand-crafted features and mistakes. We also observed that the tuning of the system and parameters relating to the size of the context-window words for fine-tuning Named Entities (w), the number of topics clusters for sensitive categories (n) and the number of neighbour words (k) for filling in the clusters can affect in a significant way the accuracy of the process results.

Finally, we can conclude that the main contribution of this work is represented by the possibility to produce annotated corpora in a semi-automatic way, where humans expertise is required only at the end of the loop, for validate and refine data automatically obtained. Preliminary results need to be improved and further experiments have to be performed, by changing the language, the size of the training and testing corpora, the tuning of the parameters in the knowledge extraction and refining step. Anyway, the preliminary results were considered promising from the domain experts that helped us in the validation stage and we are so encouraged to go further in this direction and introducing the improvements we discussed in the section before.

Acknowledgments

This work has been partially supported by MIUR - SecureOpenNets and EU SPARTA contract 830892, CyberSANE projects, and the EU project CyberSure 734815.

References

[1] Gahi, Y., Guennoun, M., Mouftah, H. T. (2016, June). Big data analytics: Security and privacy challenges. In 2016 IEEE Symposium on Computers and Communication (ISCC) (pp. 952-957). IEEE.

[2] Qiu, H., Kapusta, K., Lu, Z., Qiu, M., Memmi, G. (2019). All-or-nothing data protection for ubiquitous communication: challenges and perspectives. Information Sciences, 502, 434-445.

[3] Duan, Y., Edwards, J. S., Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. International Journal of Information Management, 48, 63-71.

[4] Voigt, P., Von dem Bussche, A. (2017). The EU general data protection regulation (GDPR). A Practical Guide, 1st Ed., Cham: Springer International Publishing.

[5] Albrecht, J. P. (2016). How the GDPR will change the world. Eur. Data Prot. L. Rev., 2, 287.

[6] Cormode, G., Srivastava, D. (2009, June). Anonymized data: generation, models, usage. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 1015-1018).

[7] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571-588.

[8] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[9] Abouelmehdi, K., Beni-Hessane, A., Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. Journal of Big Data, 5(1), 1.

[10] Acar, A., Aksu, H., Uluagac, A. S., Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys (CSUR), 51(4), 1-35.

[11] Usha, L. J., Nayahi, J. J. V. (2019). Security and Privacy in Big Data Cyber-Physical Systems. Cybersecurity and Privacy in Cyber Physical Systems, 217.

[12] Fan, W., He, J., Guo, M., Li, P., Han, Z., Wang, R. (2020). Privacy preserving classification on local differential privacy in data centers. Journal of Parallel and Distributed Computing, 135, 70-82.

[13] Pan, S. J., Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

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

[15] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[16] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[17] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

[18] Pota, M., Marulli, F., Esposito, M., De Pietro, G., Fujita, H. (2019). Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings. Knowledge-Based Systems, 164, 309-323.

[19] Aprosio, A. P., Moretti, G. (2016). Italy goes to Stanford: a collection of CoreNLP modules for Italian. arXiv preprint arXiv:1609.06204.

[20] Honnibal, M., Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear, 7(1).

[21] Syafiq, M. I., Talib, M. S., Salim, N., Haron, H., Alwee, R. (2019, August). A Concise Review of Named Entity Recognition System: Methods and Features. In IOP Conference Series: Materials Science and Engineering (Vol. 551, No. 1, p. 012052). IOP Publishing.