

Sensitive Information Detection Adopting Named Entity Recognition: A Proposed Methodology

Lelio Campanile(✉), Maria Stella de Biase, Stefano Marrone, Fiammetta Marulli, Mariapia Raimondo, and Laura Verde

Department of Matemathics and Physics, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy
{lelio.campanile,mariastella.debiase,stefano.marrone,fiammetta.marulli, mariapia.raimondo,laura.verde}@unicampania.it

Abstract. Protecting and safeguarding privacy has become increasingly important, especially in recent years. The increasing possibilities of acquiring and sharing personal information and data through digital devices and platforms, such as apps or social networks, have increased the risks of privacy breaches. In order to effectively respect and guarantee the privacy and protection of sensitive information, it is necessary to develop mechanisms capable of providing such guarantees automatically and reliably. In this paper we propose a methodology able to automatically recognize sensitive data. A Named Entity Recognition was used to identify appropriate entities. An improvement in the recognition of these entities is achieved by evaluating the words contained in an appropriate context window by assessing their similarity to words in a domain taxonomy. This, in fact, makes it possible to refine the labels of the recognized categories using a generic Named Entity Recognition. A preliminary evaluation of the reliability of the proposed approach was performed. In detail, texts of juridical documents written in Italian were analyzed.

Keywords: Named entity recognition · Sensitive information · Data privacy · Anonymization · Information extraction

1 Introduction

With the fervent expansion of Internet usage, trillions of data circulate in the virtual world every day: a critical part of this data is constituted by *sensitive information*. In particular, there are four types of sensitive data and some examples for each type are given below. *Personal Information*, including: name, surname, date of birth, address, religious orientation. *Network Identity Information* such as: IP address, MAC address or social media account. *Secret Information and Credentials*, for example: digital certificates, private key, combo login password. *Financial Information*, which includes: bank account information, credit card number, digital currency [9].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13380, pp. 377–388, 2022.
https://doi.org/10.1007/978-3-031-10542-5_26

The loss of a single sensitive data, among many, and its malicious use can be a source of damage to personal image, generation and dissemination of fake news, virtual theft to the detriment of companies and financial losses. Two are the main factors that determine the leakage of information: data structure and type of breach. It is easier to extract data, to be used in malicious way, in structured data such as databases. On the other hand, documents such as emails, text messages, social networks posts and newspaper articles[1], which are characterized by the use of Natural Language (NL), are defined as *unstructured data*. These types of documents are irregular and do not follow a predetermined pattern. Furthermore, the features of NL reside in the syntactic and semantic rules that vary according to language nationality. Although not easily processed by software, unstructured data falls within the categories of sensitive information and therefore are vulnerable to malicious attacks.

In the light of this, each country felt the need to establish data protection regulations[2]. In the People's Republic of China (PRC), the three main pillars of the personal information protection framework are the Personal Information Protection Law (PIPL), the Cybersecurity Law (CSL), and the Data Security Law (DSL). The Act on the Protection of Personal Information (APPI) regulates privacy protection issues in Japan. To safeguard privacy also in healthcare, the U.S. has adopted the Health Insurance Portability and Accountability Act (HIPAA). Also, the General Data Protection Regulation (GDPR), for European Union (EU) Member States, proposes a unified law for the processing and the circulation of sensitive data.

The GDPR allows the collected personal data to be used and processed only for the purpose stated in the consent statement [5,6]. In many cases, the purposes do not involve training a Machine Learning (ML) model. However, the principle of purpose limitation does not apply if the personal data is anonymized [4]. Several studies discussed the use of Named Entity Recognition (NER) to automatic annotate and anonymize sensitive data [14–17].

In this study, we propose a methodology that can automatically detect sensitive data from textual documents to support their obfuscation and preserve people's privacy. Linguistic processing was performed on these documents, appropriate entities were automatically recognized through the use of a NER to select the data to obfuscate. However, the recognition of such entities is not always reliable, especially for Italian. The categorization of Italian texts, in fact, is not very specific in contrast to texts written in English. Our aim is to improve the recognition of these entities based on the context domain, perfecting the sensitive information detection. Preliminary experimental tests were performed to assess the proposed methodology. A corpus of juridical documents from the Italian Court of Cassation (Corte di Cassazione) were processed.

The paper consists of the following sections: Sect. 2 presents an overview of the state of the art, the issues related to data privacy and the proposed solutions. Section 3 and Sect. 4 describe the steps of the proposed methodology and show

[1] https://www.mongodb.com/unstructured-data.
[2] https://www.dlapiperdataprotection.com/index.html.

its application to the legal domain. Section 5 explains how the methodology was evaluated and discusses the results. The article ends with an observation on the future plan.

2 Related Works

Many researchers worked to solve privacy protection problems and many techniques have been developed, such as Anonymization (not only for text [21]), Perturbation [12], Condensation [1], Randomization [3] and Fuzzy based methods [2]. The work in [20] demonstrates that the anonymization of information does not affect the *quality* of the data. In particular, anonymization can be implemented in four different ways: suppression, tagging, random substitution and generalization. The research in [13] discusses, applies and evaluates these methods to present an anonymization for documents written in Portuguese, with the aim to remove sensitive information from unstructured documents.

Suppression and tagging (both evaluated in [23], too) are used to replace the sensitive information; on the contrary, random substitution (analyzed in [10]) and generalization are used to replace sensitive information by textual expression. Their anonymization system uses the NER module to detect possible entities of relevance for the considered documents. Indeed, NER is one of the pre-processing stages of Natural Language Processing (NLP). Its goal is tagging entities in the text. There are several pre-trained libraries to recognize entities like spaCy[3] and Natural Language Processing Tool Kit (NLTK).

The concept of data quality is introduced as the problems to be settled, however, are closely linked to the purpose of using sensitive data. The systematic review in [11] classifies and discusses four categories of purpose in anonymization: data secondary usage, re-identification risk, effects on information extraction and inadequacy of current methods for heterogeneous documents. The availability of digital documents and the access to these is crucial in different fields. On the one hand, the research needs datasets to perform model training for machine learning, data exploration, qualitative and quantitave analysis. On the other hand, companies request analysis on financial, legal and clinical sensitive documents. Because of these needs, anonymization is necessary to comply with the regulations and therefore to allow a second usage of the data. In the medical field, in order to balance privacy protection and data value preserving, [19] proposes a pseudonymization approach and [8] shows that the k-anonymity algorithm degrades test coverage of Database-centric applications (DCAs).

Another observation should be made, as shown in [7]: an anonymization process can give different results depending on the application dataset. Therefore, since the data quality can invalidate the analysis, it would be necessary to carefully tailor the extraction methodology to the available data. Linked to this problem, attention can be placed on a field as delicate as the medical one: the legal field. The documents produced at each legal sentence are full of sensitive data.

[3] Here a demonstration of entities with spaCy https://explosion.ai/demos/displacy-ent.

Information extraction becomes more difficult to apply with current methodologies. In fact, although there are legal formulas recited during the sentences, the entities with the same syntactic tags have different semantic functions. Therefore, a deeper and more sensitized analysis of linguistic rules is necessary.

The approach followed in [22] uses Support Vector Machines (SVMs) to identify legal concepts and then NER to identify entities such as locations, organizations and dates. Other works use the concept of *context window*. In text sense, a context window is the set of words that compose the neighborhood of a given tagged word. The work in [27] proves that the choice of a proper fixed value for the context window size positively impacts on the overall performance of word segmentation; the same cannot be said for a higher one. This study focuses on Chinese linguistics, which is very different from the Italian and English one, and for this reason it is necessary to carry out tests to support or refute the goodness of using the context window. Here comes back the problem of tailoring the methodology on data, in particular on languages. An example is given in [18], where a pre-trained NER could not properly recognize tag in Dravidian language group, and the authors developed a NER trained on customized entities.

Considering the goals, issues and techniques analyzed in the literature, this work wants to get more than a well-performing NER customized on the rules of the Italian language in legal domain. This paper aims to focus on the information extraction phase preceding the application of anonymization. The goal is to provide a methodology disconnected from the choice of the next step to apply, be it a tagging step or a suppression one.

3 Methodology

The proposed methodology can automatically detect sensitive information in textual documents to support data obfuscation and, consequently, preserve the confidentiality and privacy of the subjects. Linguistic features are extracted from these textual documents and appropriate entities are identified. The aim is to optimize the detection of entities based on the domain of interest, in order to improve the classification of that entity as sensitive information.

As indicated in Fig. 1, to contextualize the detection of entities within a specific domain in Italian, a Domain Taxonomy (DT) is defined through an unsupervised word embedding model, aimed at assessing the words distributions, word

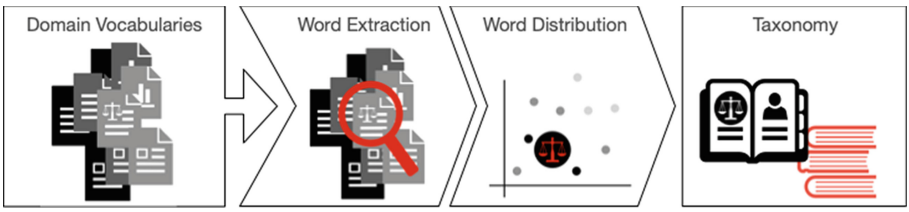

Fig. 1. Domain taxonomy definition

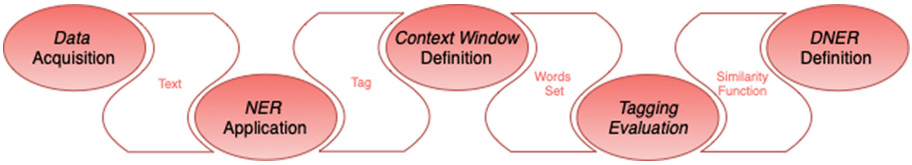

Fig. 2. Methodology phases

frequencies and words pattern co-occurrences. This taxonomy was constituted by the main entities recognized by the main NLP toolkits for Italian. For each entity are also defined subcategories that specify the domain. The main category "PER", for example, that indicates the category of people, has a series of subcategories such as, for example, minor, relative, adopted or adopter.

As shown in Fig. 2, the proposed methodology consists of several phases. Starting from the acquisition of data, a first level NER is applied, which returns labeled entities. Then, a context window is defined, in order to optimize the recognition of the obtained tags, leveraging the domain knowledge. Finally, the methodology provides the application of a specialized NER.

The first phase is data acquisition, within which textual documents to be evaluated are collected. Then, a pre-trained NER is applied to the textual documents in order to highlight entities with appropriate tags. NER is a fundamental step in NLP for the information extraction, necessary to classify entities in texts according to a list of predefined tags.

Three are the main NLP toolkits that allow to apply NER: spaCy, NLTK[4] and StanfordNLP[5]. In this work, spaCy has been adopted due to its ease of use within Python and reliability in the entities extraction [24,25]. In fact, spaCy provides pre-trained models for several languages, including Italian.

Unfortunately, this toolkit has a gap between the list of recognized tags in English and the list of possible tags for Italian. Table 1 shows the comparison between the tags provided for English[6] and Italian[7] models.

In particular, Italian pre-trained models have only four categories of possible tags (LOC, MISC, ORG and PERSON) in contrast to the eighteen categories provided by the English ones. The exiguous number of tags recognized in Italian inevitably generates a loss of specificity. In fact, a word that in English could be categorized as Geopolitical entity (GPE), in Italian would fall into the Miscellaneous entities (MISC) group. Therefore, one of the objectives to be achieved in this work is the improvement of the entities recognition for the Italian language.

To this aim, the next step of the methodology is the definition of context windows for the previously tagged words, in order to optimize the estimation of the entity tag. The choice of context is crucial to support correct entity recogni-

[4] https://www.nltk.org.
[5] https://nlp.stanford.edu/projects/project-ner.shtml.
[6] The English model *en_core_web_lg* is available at https://spacy.io/models/en.
[7] The Italian model *it_core_news_lg* is available at https://spacy.io/models/it.

tion. The context window, constructed by centering the word to be estimated, determines which contextual neighbours are considered. Several experimental tests were performed to define the suitable value for the window size, since it is fundamental to select the most appropriate size for the context window.

Table 1. spaCy tag comparison between English and Italian

Tag	Description	English	Italian
CARDINAL	Numerals that do not fall under another type	✓	✗
DATE	Absolute or relative dates or periods	✓	✗
EVENT	Named hurricanes, battles, wars, sport events, etc.	✓	✗
FAC	Buildings, airports, highways, bridges etc.	✓	✗
GPE	Countries, cities, states	✓	✗
LANGUAGE	Any named language	✓	✗
LAW	Named document made into law	✓	✗
LOC	Non-GPE locations	✓	✓
MISC	Miscellaneous entities	✗	✓
MONEY	Monetary values, including unit	✓	✗
NORP	Nationalities or religious or political groups	✓	✗
ORDINAL	“first”,“second”, etc.	✓	✗
ORG	Companies, agencies, institutions, etc.	✓	✓
PERCENT	Percentage, including “%”	✓	✗
PERSON	People, including fictional	✓	✓
PRODUCT	Objects, vehicles, foods etc.	✓	✗
QUANTITY	Measurements, as of weight or distance	✓	✗
TIME	Times smaller than a day	✓	✗
WORK_OF_ART	title of book, songs, etc.	✓	✗

The successive phase of the methodology consists in comparing the labeled entities obtained from the application of the NER with the entities defined in DT.

The accuracy of each tag was evaluated through a *similarity function*. In this study, the similarity function D has been defined taking into account lexical, semantic and taxonomy measures. Words that have a similar characters sequence are lexically similar. On the other hand, if the words are lexically different, but they express the same concept, then they are semantically similar. Additionally, it is possible to obtain another measure evaluating the semantic similarity of a given word in comparison to the words in the DT.

Several methods can be used to calculate the lexical, semantic and taxonomy similarities. Appropriate metrics and algorithms, such as for example Levenshtein distance, Latent Semantic Analysis (LSA) or Lesk similarity, are computed to estimate the grade of similarity [26]. The function D is defined as follows:

$$D(w_c, w_{DT}) = p_L * L + p_C * C + p_K * K \tag{1}$$

with

$$L = \sum_{i=1}^{n_l} l_i(w_c, w_{DT}) \qquad C = \sum_{i=1}^{n_c} c_i(w_c, w_{DT}) \qquad K = \sum_{i=1}^{n_k} k_i(w_c, w_{DT}) \tag{2}$$

where w_c and w_{DT} indicate, respectively, the words of the context window and the words of the DT, while L, C and K are the sum of similarity metrics (l_i, c_i and k_i), respectively, for the lexical, semantic and taxonomy similarities, and p_L, p_C and p_K are the corresponding weights for the three measures.

A linear constraint placed on the definition of the function is that each weight takes on a value in the range $[0, 1]$. The similarity function D is applied to each word w_c in the context window to evaluate its similarity to the words of the DT. Therefore, assuming that the DT has cardinality M, a value vector v of size M is obtained for each word of the context window, containing the results of the function D. It is worth reminding that the DT consists of different N categories and that each word of the DT belongs to exactly one category.

To associate a category to each w_c, the average of all calculated distances between the w_c among the same category of the DT is computed, obtaining a vector $\tilde{v}$ of size N, i.e., the number of categories. Each number of this vector indicates how much the w_c is "near" to the category.

In order to decide the category, the N values of the vector are compared to a threshold value, α. This value has been found experimentally and a rule-based approach was applied to associate w_c to a specific category, as shown in Algorithm 1. Each value $\tilde{v}_i$ in $\tilde{v}$ is compared to the threshold. If only one value exceeds α, then the corresponding category is associated to the w_c. If several values exceed the threshold, the category corresponding to the largest $\tilde{v}_i$ is chosen. By doing so, each w_c with a specific associated category becomes a sentinel word w_s.

Algorithm 1. The rule-based adopted approach.

Require: $\alpha > 0$
  if $v_i \geq \alpha$ **then**
    $w_c \leftarrow w_s$ ▷ Algorithm for the definition of w_s.
  else
    $w_c \leftarrow Not \quad Considered$
  end if

The sentinel word, which was assigned a specific DT-subcategory, will improve the classification of w_t. But several w_s could be included in the considered context window. In this case, to support the definition of the sub-category of the w_T, the DT-subcategory of w_s that achieved the best $\tilde{v}_i$ and had the same category of the w_T was chosen.

Finally, the methodology provides the application of a Domain NER (DNER), which consists in the definition of a specialized context window-based NER. Such a DNER is applied to selected textual documents to perform a second tagging evaluation to assess the reliability of the methodology.

4 Case Study

In order to evaluate the reliability of the proposed approach, preliminary experimental tests were performed. Appropriate documents concerning legal domain were considered. In detail, recent sentences selected from archive of the Italian Court of Cassation (Corte di Cassazione) were processed.

The dataset was composed by 29 civil judgments on divorce, separation or succession of property. All documents are written in Italian. The average number of words per document is 896, whilst the total number of words in the whole dataset is 26001.

An open-source NLP toolkit was used in the first step of our experiments. In detail, spaCy pre-trained model[8] for Italian was adopted. This model was trained by using data selected from four database: UD Italian ISDT v2.8[9], an Italian corpus annotated according to the Universal Dependencies (UD) annotation, obtained by the conversion from Italian Stanford Dependency Treebank (ISDT); WikiNER[10], a learning multilingual named entity recognition from Wikipedia; Lemmatization Lists[11], a dataset of several lemma/token pairs; and Explosion fastText Vectors[12]. As said in Sect. 3, for the Italian language, this tool returns data annotated with the following Named Entity (NE) classes: LOC, PER, ORG, MISC.

A DT was defined for this juridical sector, according to the approach described in [14]. For each one of the main categories, several subcategories were identified. As an example, Table 2 shows the subcategories of the PER category in the considered domain.

For each subcategory, a number of useful words were collected for the recognition of the subcategory itself. Taking into account "AVVOCATO" (i.e., lawyer), Table 3 shows some of the words gathered to identify this subcategory, like "Avv.", "Avvocato", "Difensore".

The application of the pre-trained NER to the legal documents considered led to the identification of 2378 labels. An average of 82 labels were identified for each document. We focused on the PER category because it was the predominant one in our documents, thus it was the best one to work on to check the effective optimization of the labeling.

[8] https://spacy.io/models/it#it_core_news_lg.
[9] https://github.com/UniversalDependencies/UD_Italian-ISDT.
[10] https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500.
[11] https://github.com/michmech/lemmatization-lists/.
[12] https://spacy.io/.

Table 2. PER category

Category	Subcategory
PER	GIUDICE
	INABILE
	RAPPRESENTANTE INCAPACI
	RAPPRESENTANTE CAPACI
	INCAPACE
	MINORENNE
	PARENTE
	PARTE
	ADOTTATO
	ADOTTANTE
	AVVOCATO

Table 3. Recognition of AVVOCATO subcategory

Subcategory	Word
AVVOCATO	Avv.
	Avvocato
	difensore
	difeso
	legale
	patrocinante
	patrocinato
	patrocinio
	rappresentante
	rappresentato
	patrocinatore

As indicated in Table 4, the amount of recognized words, in all documents, labeled as PER category was 957, each document has an average of 33 PER labels.

For each document a version annotated by domain experts was available to evaluate the correctness of the identified categories. For the annotated documents, 1715 entities were individuated, an average of 59 entities were recognized for each document. Relating to the PER category, 814 were the labeled entities, with an average of 28 per document.

An opportune context window was chosen. The size of the window used in the analysis was 5, which has been found by experimental tests. By choosing this size, the results obtained from the application of the DNER were of 86 words for which a specific subcategory was identified and used as label. This means that about a 10% of the PER labels could be specialized to a specific subcategory.

Table 4. Case study: results achieved for the recognition of PER category.

	Total PER words	Average per document
NER	957	33
Domain expert	814	28

5 Conclusions

Nowadays, the increasing opportunities to capture and share sensitive information and data have increased the risks of privacy violations. Therefore, techniques and countermeasures to preserve and protect sensitive data of people are necessary. Data anonymization can play a critical role to protect privacy of people obscuring sensitive information to avoid the identification of the subject.

This paper presents a methodology to automatically anonymize personal data. This methodology is able to extract linguistic features, from textual documents, necessary to identify opportune entities fundamental to recognize personal data. In detail, an automatic NER was used to detect the main entities in these documents. Then, an opportune domain taxonomy has been defined and used to improve the categorization of these entities.

Preliminary experimental tests were performed to evaluate the reliability of the proposed approach. Juridical documents from the Italian Court of Cassation (Corte di Cassazione) were processed. Appropriate entities were extracted by using the pre-trained NER provided by spaCy, an open-source NLP toolkit. An opportune DT was defined to improve the recognition of categories.

Our future plans will include an improvement of the proposed methodology considering other similarity metrics. Additionally, this is a preliminary study, in which only some aspects of the proposed approach were evaluated. Further aspects of the proposed methodology will be analyzed, not only in the juridical domain, but even in other sector such as healthcare or financial.

Acknowledgement. The work of Maria Stella de Biase is granted by PON Ricerca e Innovazione 2014/2020 MUR—Ministero dell'Università e della Ricerca (Italy)—with the Ph.D. program XXXVI cycle. The work of Mariapia Raimondo is granted by INPS—Istituto Nazionale di Previdenza Sociale (Italy)—with the PhD program XXXVI cycle.

The work of Fiammetta Marulli is granted by the project "Attrazione e Mobilità dei Ricercatori" Italian PON Program (PON_AIM 2018 num. AIM1878214-2). The work of Laura Verde is granted by the "Predictive Maintenance Multidominio (Multidomain predictive maintenance)" project, PON "Ricerca e Innovazione" 2014–2020, Asse IV "Istruzione e ricerca per il recupero"-Azione IV.4-"Dottorati e contratti di ricerca su tematiche dell'innovazione" programme CUP: B61B21005470007.

The research leading to these results has received funding from the projects "ANDROIDS" and "WAIILD TROLS" funded, respectively, by the programs V:ALERE 2019 and V:ALERE 2020 of the Università della Campania "Luigi Vanvitelli".

References

1. Aleroud, A., Yang, F., Pallaprolu, S.C., Chen, Z., Karabatis, G.: Anonymization of network traces data through condensation-based differential privacy. Digit. Threat. Res. Pract. **2**(4), 1–23 (2021). https://doi.org/10.1145/3425401
2. Ashok, V.: Fuzzy-based methods for privacy-preserving data mining, pp. 348–353 (April 2011). https://doi.org/10.1109/ITNG.2011.68
3. Batmaz, Z., Polat, H.: Randomization-based privacy-preserving frameworks for collaborative filtering. Procedia Comput. Sci. **96**, 33–42 (2016). https://doi.org/10.1016/j.procs.2016.08.091
4. Biesner, D., et al.: Anonymization of German financial documents using neural network-based language models with contextual word representations. Int. J. Data Sci. Anal. **13**(2), 151–161 (2022)
5. Campanile, L., Iacono, M., Levis, A.H., Marulli, F., Mastroianni, M.: Privacy regulations, smart roads, blockchain, and liability insurance: putting technologies to work. IEEE Secur. Priv. **19**(1), 34–43 (2020)
6. Campanile, L., Iacono, M., Marulli, F., Mastroianni, M.: Privacy regulations challenges on data-centric and IoT systems: a case study for smart vehicles. In: IoTBDS, pp. 507–518 (2020)
7. Ferrández, Ó., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M.: Generalizability and comparison of automatic clinical text de-identification methods and resources. In: AMIA Annual Symposium Proceedings, vol. 2012, p. 199. American Medical Informatics Association (2012)
8. Grechanik, M., Csallner, C., Fu, C., Xie, Q.: Is data privacy always good for software testing? pp. 368–377 (November 2010). https://doi.org/10.1109/ISSRE.2010.13
9. Guo, Y., Liu, J., Tang, W., Huang, C.: Exsense: extract sensitive information from unstructured data. Comput. Secur. **102**, 102156 (2021). https://doi.org/10.1016/j.cose.2020.102156. https://www.sciencedirect.com/science/article/pii/S0167404820304296
10. Kang, J.S.: An improvement of privacy-preserving scheme based on random substitutions (March 2022)
11. Langarizadeh, M., Orooji, A., Sheikhtaheri, A.: Effectiveness of anonymization methods in preserving patients' privacy: a systematic literature review. Stud. Health Technol. Inf. **248**, 80–87 (2018)
12. Liu, L., Thuraisingham, B.: The applicability of the perturbation model-based privacy preserving data mining for real-world data, pp. 507–512 (January 2007). https://doi.org/10.1109/ICDMW.2006.155
13. Mamede, N., Baptista, J., Dias, F.: Automated anonymization of text documents, pp. 1287–1294 (July 2016). https://doi.org/10.1109/CEC.2016.7743936
14. Martinelli, F., Marulli, F., Mercaldo, F., Marrone, S., Santone, A.: Enhanced privacy and data protection using natural language processing and artificial intelligence. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
15. Di Martino, B., Marulli, F., Graziano, M., Lupi, P.: PrettyTags: an open-source tool for easy and customizable textual multilevel semantic annotations. In: Barolli, L., Yim, K., Enokido, T. (eds.) CISIS 2021. LNNS, vol. 278, pp. 636–645. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79725-6_64

16. Di Martino, B., Marulli, F., Lupi, P., Cataldi, A.: A machine learning based methodology for automatic annotation and anonymisation of privacy-related items in textual documents for justice domain. In: Barolli, L., Poniszewska-Maranda, A., Enokido, T. (eds.) CISIS 2020. AISC, vol. 1194, pp. 530–539. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-50454-0_55
17. Marulli, F., Verde, L., Marrone, S., Barone, R., De Biase, M.S.: Evaluating efficiency and effectiveness of federated learning approaches in knowledge extraction tasks. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2021)
18. Melinamath, B.: Rule based methodology for recognition of Kannada named entities. Int. J. Latest Trends Eng. Technol. (IJLTET) **3**, 50–58 (2014)
19. Neubauer, T., Riedl, B.: Improving patients privacy with pseudonymization. Stud. Health Technol. Inf. **136**, 691–696 (2008). https://doi.org/10.3233/978-1-58603-864-9-691
20. Panackal, J., Pillai, A.: Adaptive utility-based anonymization model: performance evaluation on big data sets. Procedia Comput. Sci. **50**, 347–352 (2015). https://doi.org/10.1016/j.procs.2015.04.037
21. Perero-Codosero, J., Espinoza-Cuadros, F.M., Gómez, L.: X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. Comput. Speech Langu. **74**, 101351 (2022). https://doi.org/10.1016/j.csl.2022.101351
22. Quaresma, P., Gonçalves, T.: Using linguistic information and machine learning techniques to identify entities from juridical documents. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 44–59. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12837-0_3
23. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**, 1010–1027 (2001). https://doi.org/10.1109/69.971193
24. Schmitt, X., Kubler, S., Robert, J., Papadakis, M., LeTraon, Y.: A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In: 2019 6th International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 338–343. IEEE (2019)
25. Spring, R., Johnson, M.: The possibility of improving automated calculation of measures of lexical richness for EFL writing: a comparison of the LCA, NLTK and SpaCy tools. System **106**, 102770 (2022)
26. Vijaymeena, M., Kavitha, K.: A survey on similarity measures in text mining. Mach. Learn. Appl. Int. J. **3**(2), 19–28 (2016)
27. Xu, W., Nong, G.: A study for extracting keywords from data with deep learning and suffix array. Multimedia Tools Appl. **81**(5), 7419–7437 (2022). https://doi.org/10.1007/s11042-021-11762-7