Journal of Intelligent Information Systems (2023) 61:541–568
https://doi.org/10.1007/s10844-023-00782-4

An AI framework to support decisions on GDPR compliance

Filippo Lorè[1] · Pierpaolo Basile[1] · Annalisa Appice[1,2] · Marco de Gemmis[1] · Donato Malerba[1,2] · Giovanni Semeraro[1]

Received: 23 November 2022 / Revised: 13 February 2023 / Accepted: 14 February 2023 /
Published online: 18 March 2023

Abstract
The Italian Public Administration (PA) relies on costly manual analyses to ensure the GDPR compliance of public documents and secure personal data. Despite recent advances in Artificial Intelligence (AI) have benefited many legal fields, the automation of workflows for data protection of public documents is still only marginally affected. The main aim of this work is to design a framework that can be effectively adopted to check whether PA documents written in Italian meet the GDPR requirements. The main outcome of our interdisciplinary research is INTREPID (artIficial iNTelligence for gdpR compliancE of Public admInistration Documents), an AI-based framework that can help the Italian PA to ensure GDPR compliance of public documents. INTREPID is realized by tuning some linguistic resources for Italian language processing (i.e. SpaCy and Tint) to the GDPR intelligence. In addition, we set the foundations for a text classification methodology to recognise the public documents published by the Italian PA, which perform data breaches. We show the effectiveness of the framework over a text corpus of public documents that were published online by the Italian PA. We also perform an inter-annotator study and analyse the agreement of the annotation predictions of the proposed methodology with the annotations by domain experts. Finally, we evaluate the accuracy of the proposed text classification model in detecting breaches of security.

Keywords GDPR · Italian language processing · Classification · Text data engineering

1 Introduction

In 2018, the European Union (EU) introduced the General Data Protection Regulation 2016/679 (GDPR)[1] to update and unify the data protection regulation across the EU states,

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj

✉ Annalisa Appice
annalisa.appice@uniba.it

Extended author information available on the last page of the article.

so that each member state no longer needed to write its own data protection laws and laws were consistent across the entire EU (Hoofnagle et al., 2019). From that moment on, the GDPR represented the primary legislation regulating how companies that, regardless of their location, marketed goods or services to the EU citizens, had to protect EU citizens' personal data, i.e. any data that relate to an identified or identifiable EU living person. Nowadays the GDPR, with its 11 chapters and 99 articles, mandates a baseline set of privacy and data protection standards for better safeguarding the processing and movement of EU citizens' personal data. Some of the standards defined with the GDPR include: requiring the consent of subjects for data processing, anonymizing collected data to protect privacy, providing data breach notifications, safely handling the transfer of data across borders and requiring certain companies to appoint a data protection officer to oversee GDPR compliance (Savic & Veinovic, 2018). In addition, the GDPR encourages compliance allowing each EU member state's Data Protection Authority (DPA) – the independent public authority that supervises the GDPR application – to fine violators stiff penalties. For example, organizations that fail to comply with certain GDPR standards may be fined up to 2% or 4% of total global annual turnover or 10 million € or 20 million €, whichever is greater (Savic & Veinovic, 2018).

Also the Public Administration (PA) must respect the key principles of the GDPR by guaranteeing fair and lawful processing, purpose limitation, data minimisation and data retention when processing personal data relating to EU citizens (Blume, 2016; Ricci, 2018). In particular, the PA is required to appoint a Data Protection Officer (DPO) who ensures that appropriate technical and organisational measures are implemented to secure personal data. Whenever personal data are disclosed accidentally or unlawfully to unauthorised recipients or are temporarily unavailable or altered, the breach must be notified to the DPA without undue delay and at the latest within 72 hours after having become aware of the breach. The PA may also need to inform individuals about the breach.

The DPA can take a range of actions in cases of GDPR non-compliance in the PA. In the case of a likely infringement, a warning may be issued. In the case of an infringement, the possibilities include a reprimand or a temporary or definitive ban on the processing. In some countries, such as Italy,[2] public bodies may also be subject to administrative fines (Mc Cullagh et al., 2019). For example, on July 20, 2021, the Italian DPA fined Trento health authority €150,000 for unlawful disclosure of patient health data.[3] On the other hand, individuals can also claim a compensation where a public document is in breach of the GDPR and they have suffered material damages (e.g. financial loss) or non-material damages (e.g. reputational loss or psychological distress).

Despite technologies in document management and workflow automation are developing with unprecedented speed, the PA still entrusts the GDPR compliance mainly to human operators who may slow down processes and lack of adequate education in the regulation (Di Nicola et al., 2016). On the other hand, efficiency in recognising and reporting breaches of security is a requirement coherent with the EU vision to make the PA open, efficient, inclusive, borderless and user-friendly. In this scenario, Artificial Intelligence (AI) techniques can cover a key role by providing a new digital environment for public services also referred to data protection. Following this intuition, Kingston (Kingston, 2017) has recently promoted the investigation of AI techniques for different GDPR relevant tasks such as:

[2]Norms contained in the Italian Personal Data Protection Code (Legislative Decree 196/2003) were aligned with provisions introduced by GDPR with the legislative decree n. 101/2018 published in the Official Gazette n. 205 on September 4, 2018.

[3]https://www.dataguidance.com/news/italy-garante-fines-trento-health-authori ty-150000

following compliance checklists and codes of conduct, supporting risk assessments, complying with the regulations regarding technologies that perform automatic profiling, and complying with new regulations concerning recognising and reporting breaches of security. The data protection issues introduced by GDPR have an impact also on document processing systems. In particular, we focus on the implications of the GDPR on the publication of unstructured (textual) documents which might disclose personal information. In this context, AI techniques such as Natural Language Processing (NLP) can help to automatically detect those parts in textual documents that might constitute a data breach. NLP is a subarea of AI that processes natural human language, either in text or voice. Google's search suggestions or spell checkers are familiar examples of these techniques. NLP is nowadays largely used in several intelligent tools available to lawyers. For instance, one of the most popular and helpful applications of NLP in law is in legal research: NLP-powered legal search engines can look for concepts, not just specific keywords, helping lawyers find what they need faster. The recognition of text patterns in documents is useful to detect relevant text fragments that might contain personal information that must be protected. Machine Learning techniques, such as text categorization, are also exploited in this context to build models that predict whether personal data are disclosed in a document, by learning from previous examples of data breach.

This paper takes up the challenge and investigates how AI techniques can actually help to automate (or semi-automate) the data protection workflows of the PA, in order to reduce risks of breaches of security in public documents. This is investigated in a case study involving the Italian PA that, with 60 million people, 8000 municipalities, and 22,000 local administrations, has actively embraced the digital transformation of public administration efficiencies (Datta, 2020). The main contributions of this work are:

1. We focus on the problem of detecting breaches of security related to unlawful disclosure of health information in public documents, according to the GDPR standards, which is one of the main causes of administrative fines. Specifically, we consider the problem as a binary classification task and we reduced the field of investigation to documents of the Italian public administration. The main outcome of our investigation is an AI framework – INTREPID (artIficial iNTelligence for gdpR compliancE of Public admInistration Documents) – that is designed to process the public documents produced by the Italian PA and classify them as compliant or non-compliant with the GDPR standards.
2. Given the lack of benchmarks for the GDPR compliance analysis of public documents and the need for building a suitable training set, we collected a corpus of public documents that were published online from various municipalities of the Italian PA. The corpus contains both 45 GDPR-compliant documents that correctly protect personal data and 45 non GDPR-compliant documents that unlawfully disclose health information. Documents in the corpus were annotated by two experts to identify named entities as described in Section 3. Starting from these annotations, we developed an automatic pipeline for anonymizing the documents that were finally used for the evaluation of the performance of the proposed framework. In addition, we focused on an information extraction task from text written in Italian (Attardi et al., 2015; De Felice et al., 2018). Therefore, we resort to specific linguistic resources developed for Italian language processing, but tune them to the GDPR intelligence. Despite this step is language dependent, it can be generalized to other languages with little effort. The collected corpus of anonymised PA documents written in Italian was essential to perform

an experimental analysis of INTREPID framework that demonstrated the effectiveness of the framework.

3. We address crucial aspects that typically arise in AI models, namely feature extraction and classification. Although various approaches have been proposed for both tasks, no related literature has been published to date related to their specific application for the data protection in PA documents written in Italian. In this regard, we investigate how to perform an information extraction step that seeks to locate and classify named entities in Italian text into pre-defined categories such as the names of persons, organizations, locations, health status, administrations and expressions of *"omissis"*. Furthermore, we provide insights in how setting up text classification on both bag-of-words features and named entity-based features.

The remainder of the paper is organised as follows. Section 2 overviews works prompting this investigation of AI for ensuring data protection in the Italian PA, as well as the background of AI techniques in text classification problems. Section 3 describes the corpus of the PA documents written in Italian we have collected for this study, while Section 4 presents our proposed framework for verifying the GDPR-compliance of Italian PA documents. Experimental evaluation methodology and results are presented in Section 5. Section 6 illustrates benefits of the proposal and how it could be enhanced in the future research. Finally, Section 7 concludes the paper.

2 Related work

Our work reflects the recent interest in the role of AI techniques for text data in a variety of law domains (Dadgostari et al., 2020; De Martino et al., 2022; Tagarelli & Simeri, 2021), including the GDPR domain (Kingston, 2017). In this regard, we first overview recent research works that have mainly contributed to the investigation of AI in the GDPR application. Then, as our focus is the GDPR compliance of the text documents produced by the Italian PA, we describe some works that adopt NLP for the identification and extraction of sensitive data, especially in PA documents.

2.1 AI and GDPR

AI used in accordance with the provisions of EU and national legislation is expected to significantly increase the efficiency level of both companies and public offices (Stamova & Draganov, 2020). This expectation has prompted several recent studies to explore the relation between AI and GDPR. These studies can be mainly grouped in two fields: GDPR compliance of AI and AI for GDPR compliance.

2.1.1 GDPR compliance of AI

The problem of the GDPR compliance of AI was officially formalised in early 2020, when the EU Commission published a White Paper on AI regulation.[4] This paper highlighted the need to review the EU's legislative framework with a view to making it fit for the current technological developments. In particular, it clarifies that the GDPR always applies itself to

[4]https://ec.europa.eu/info/sites/default/files/commission-white-paper-artific ial-intelligence-feb2020_en.pd (last access: 2021/10/13)

AI when AI techniques process personal data, perform profiling, as well as make automated decisions based on personal data and/or that affect the data subjects. In response to these issues, the proposal for a EU AI Regulation,[5] published in April 2021, is the latest addition to the EU Commission's Digital Strategy, which takes the first steps in the GDPR regulation of AI. In this regard, the extent to which AI fits into the GDPR conceptual framework has recently been explored by Sartor and Lagioia (2020). This study delineates the legal bases for AI applications to personal data and the duties of information concerning AI systems, especially those involving profiling and automated decision-making.

In this perspective, also the "right to explanation" formulated by the GDPR poses a non-trivial technical challenge to harness the full power of machine learning or AI systems while operating with logic interpretable to humans (Selbst & Powles, 2017). In fact, the GDPR creates rights to "meaningful information about the logic involved" when an individual is subject to a decision based solely on automated processing that significantly affects him or her. For instance, doctors using AI systems to propose treatment plans need to know why a certain course of action has been identified in order to explain the decision to patients. There is a need for methods to certify, explain, and audit inscrutable systems. The debate around the "right to explanation" has drawn high interest both from the law and the AI community, and the question of whether it is technically possible is still an open issue. DARPA, the Defense Advanced Research Projects Agency, launched an "Explainable AI" initiative in 2016, with the aim of producing a toolkit library consisting of machine learning and human-computer interface software modules that could be used to develop future explainable AI systems. Following this research direction, the study of Sovrano et al. (2020) has recently introduced a user-centred explanatory model for Trustworthy AI, compliant with GDPR. In particular, the study introduces a definition of user-centred explanations as Explanatory Narratives, based on concepts drawn from ISO 9241, and presents a formal model of an Interactive Explanatory Process by identifying both the fundamental properties of a good explanation and the heuristics for the exploration of the explanatory space.

Instead the study of Meszaros and Ho (2021) identifies the differences between how academic and commercial research apply the GDPR in the development of AI products and services. The main outcome is that companies conduct commercial research that might not have in place a similar level of ethical and institutional safeguards as academic researchers. In addition, the study stresses the need to find the proper balance between privacy and innovation. In general, the EU vision is that transparency and accountability can build trust together within the GDPR compliance of AI. However, this is nowadays an open challenge that still requires further regulations in terms of EU laws, and new scientific and technological effort in terms of responsible developments.

2.1.2 AI for GDPR compliance

As the GDPR became law, a last-minute rush started to become compliant. Many companies started to offer advice, checklists and consultancy on how to comply with the GDPR. In such an environment, AI emerged as a key technology by providing best advice, asking all and only the relevant questions and carrying out assessments (Kingston, 2017). This thought was inspired by a few studies that, even before the GDPR became effective on May 2018, started exploring how legal knowledge-based systems (e.g. rule-based systems) may be used

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (last access: 2021/10/13)

as intelligent checklists to verify the GDPR compliance in the risk analysis (Al-Abdulkarim et al., 2016; Kingston, 2017; van Engers, 2005).

Inspired by these studies, AI techniques are explored for the GDPR compliance in the automation of the legal evaluation of privacy policies, which are the primary channel through which companies inform users about their data collection and sharing practices. Since the seminal AI studies (Contissa et al., 2018; Sánchez et al., 2021) in this domain rely on datasets for classification manually annotated by experts, automatic annotation has started receiving some attention. In this regard, a deep learning approach is explored by Harkous et al. (2018), in order to automatically annotate previously unseen privacy policies with high-level and fine-grained labels from a pre-specified taxonomy.

On the other hand, recent developments in NLP have led to an increased interest in the exploration of NLP approaches for text processing in various data protection problems, e.g. text anonymization, in contexts where text data (e.g. police reports, healthcare dossiers) are either sensitive by nature or protected by privacy laws (e.g. GDPR). Mozes and Kleinberg (2021) has recently revised recent studies that leverage NLP in text anonymization by analysing the evaluation criteria to assess the effective ability to protect individuals from being re-identified with AI-based methods.

AI for GDPR compliance also attracts attention in process mining. This is a field where AI techniques are commonly leveraged to provide insights on business processes by relying mainly on process execution data (van der Aalst, 2016). In process mining, the problem of how enabling the GDPR compliance of business processes is originally formulated by Zaman et al. (2019). Following the guidelines defined in this preliminary study, process mining techniques have recently adapted for ascertaining compliance of business process execution to data subject rights and enabling GDPR-compliant business process discovery (Zaman & Hassani, 2020).

Finally, a semantic model to represent GDPR consents is studied by Davari and Bertino (2019). This model is explicit, understandable and reusable. In addition, it is coupled to a blockchain-based model for ensuring that organisations comply with GDPR with respect to user consents.

2.2 NLP for the recognition of sensitive data

When developing an automatic system to classify documents as compliant with GDPR or not, one of the problems we had to face with is the extraction of relevant entities and meaningful features. The recognition of named entities and the discovery of relations among them are the core tasks of the proposed framework. In fact, many GDPR violations in Italian Public Administration domain are represented by publication of documents containing personal data (for instance the name of an employee suffering from a certain disease). Therefore, we focus on the definition of an appropriate Named Entity Recognition (NER) strategy for the above mentioned framework. In particular, we evaluate the use of NER as a way to detect personally identifiable information within textual documents. NER is one of NLP tasks that aims to find and classify named entities present in a text into specific and pre-defined categories (Yadav & Bethard, 2019).

Previous research has adopted NER to identify entities which can be helpful for text mining and categorisation in public administration documents. For instance, (Romano et al., 2020) propose a framework for the extraction of data from sentences issued by the Court of Cassation (last degree of judgement in the Italian judicial system). In particular, the framework is based on NER for the detection of the names of companies and their legal

form; recognised entities are then linked with additional information from business registers, creating the conditions for the analysis of criminal events.

Another work which has a relation with the proposed framework is Silva et al. (2020), in which the authors evaluate the use of NER as a way to identify, monitor and validate Personally Identifiable Information (PII) in contracts. In particular, they evaluate the performance of two tools (Stanford CoreNLP and SpaCy) and show how NER is actually effective for the automated monitoring of PII in different scenarios.

The work described in Di Cerbo and Trabelsi (2018) highlights the limitation of deterministic approaches, such as regular expressions, to detect personal information into semi-structured or non-structured archives. Consequently, NER techniques based on Naive Bayes and Convolutional Neural Networks are adopted to identify the presence and the nature of personal information into data coming from social media.

Another recent work focused on NER for the sensitive data covered by GDPR is Dias et al. (2020). The authors propose a hybrid approach to the problem of NER for the Portuguese language that combines several techniques such as rule-based, lexical-based models, machine learning algorithms, and neural networks. The use of different methodologies covered all classes of entities that represent sensitive data. This work underlines again that named entities play an important role in administrative acts, in particular for the recognition of sensitive data, and that an effort is required to adapt state-of-art algorithms in this specific context. For instance, existing Italian corpora annotated with named entities are not optimal for training a NER for the PA domain, due to the differences between "bureaucratic" language and standard Italian, such as the use of uncommon, formal terms or specific abbreviations.

An attempt to face with this problem is made by Passaro et al. (2017), in which the authors describe the process of designing a NER system for Italian PA documents. They create a new corpus from scratch starting from administrative documents published by municipalities and then adapt a general NE recognizer to extend the standard NE classes to other entity types particularly relevant in the context of municipalities. We adopt a similar approach to recognise entity types which could be relevant for the GDPR-compliance. In addition, we study how the recognised named entities can be injected into a text data engineering step, in order to train a classification model able to analyse the GDPR-compliance of Italian PA documents.

Finally, NER-based approaches are widely explored in text anonymization problems. For example, Adams et al., (2019) describes a system that integrates both a NER module and a co-reference module. This system allows us to identify chunks of human-computer dialogue texts, which contain sensitive tokens in specific categories (e.g. personal names, addresses, facilities, organisations). Francopoulo and Schaub (2020) adopts a NER module that cascades pattern matching rules to recognise entities in several categories (e.g. personal names, locations, companies, email addresses) and performs anonymization operations in the context of Customer Relationship Management. Biesner et al. (2022) has recently explored the performance of NLP methods based on recurrent neural nets and transformer architectures to detect and anonymize sensitive information in German financial and legal documents.

Finally, Csányi et al. (2021) explores the performance of NER-based tools in anonymization of legal documents of Hungarian courts by drawing the conclusion that mathematical statistical analysis is crucial to filter unique events that may serve as primary identifiers (e.g. the surgeon amputates the wrong leg). However, this study also highlights the need of using machine learning-based methods together with anonymization models to reduce re-identification risk. Our work follows this research direction by combining NLP methods for NER and machine learning methods for classification.

2.3 Novel contributions

Our study can be classified under the umbrella of AI for GDPR compliance. Similarly to the majority of studies under this umbrella, we resort to a text classification formulation of the problem and adopt NLP techniques, and specifically NER techniques, to extract useful information from unstructured text and classify the text into pre-defined categories. At present, NER tools have been already studied for data protection and text anonymization in various contexts (e.g. chat texts, legal documents). So, a novelty of our study is the specific GDPR compliance problem addressed using AI. In fact, to the best of our knowledge, this is the first work that investigates how an AI framework can be effectively used to automate the GDPR intelligence involved in the data protection of the text corpora of the Italian PA.

Notably, a major difficulty addressed in the beginning of this study was the absence of benchmark data. This required the preparation of a corpus and the use of an appropriate pipeline to balance the need of replacing any identified or identifiable information with artificial identifiers and the fact that the GDPR check does not apply to anonymized information.

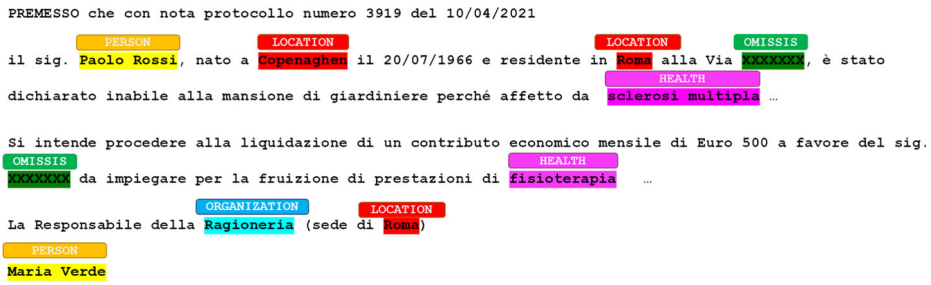
Another difference with respect to previous studies concerns the work done to tune specific linguistic resources for Italian language processing to the GDPR intelligence of PA documents. A tuning work is also performed in Passaro et al. (2017), but without the exploration of the performance of the machine learning algorithms for the document classification step. In general, previous studies focus on data protection through NER investigation. Instead, in our study, we also explore the performance of various classification algorithms (i.e., Support Vector Machine, Random Forest and XGBoost) trained on both Bag-of-Word (boW) and NER-based engineered data. To this purpose, we define several NER-based text engineered schemes and evaluate their performance in combination with both BoW information and various classification algorithms.

Notably, classification models are also trained in Contissa et al. (2018) and Sánchez et al. (2021) for classifying privacy policies of companies according to their compliance (or not) with the data protection goals of the GDPR. However, both studies train Support Vector Machines from data sets manually annotated by experts. Differently, we address a more complex learning problem, where annotations must be performed automatically with a NER tool tuned for the specific problem under study.

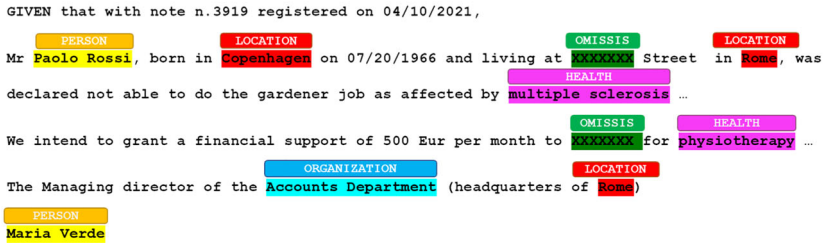
3 Corpus preparation

We prepared an Italian corpus of 90 text public documents that were published on the websites of some Italian Municipalities (74 documents) or Public Hospitals (16 documents). The collected documents are deliberations or determinations referred to work-related diseases (30 documents), welfare payments (30 documents) and financial support for health (30 documents). These documents were published by PA according to the Decree-Law No. 33/2013 of the Italian Law. The GDPR risk assessment conducted on the corpus revealed that within 45 out of 90 documents, GDPR data breaches were found.

The collection of these documents was performed according to the Decree-Law No. 33/2013 that in article 3 establishes that «All documents, information and data subject to civic access, including those subject to mandatory publication pursuant to current legislation, are public and anyone has the right to know them, to use them free of charge, and to use and reuse them».



(a) Named entity annotation (Italian version)



(b) Named entity annotation (English version)

Fig. 1 Named entities annotated in a fictional text: text written in Italian (Fig. 1a) and English translation (Fig. 1b)

The corpus was annotated by two independent annotators, experienced GDPR lawyers, who read the text thoroughly, located the target entities, highlighted them on the annotation platform,[6] and chose categories from a predetermined list: *administration, health, location, omissis, organisation* and *person*. In addition to the standard categories (*person, location, organisation*), we introduced some domain-specific categories, to recognise roles in public administration and disease names, which may be relevant for identifying GDPR non-compliant documents. The domain-specific categories were suggested by the annotators. After reading all documents in the corpus, they hypothesized categories of potentially relevant entities. An example of annotated fictional text is reported in Fig. 1.

To assess the quality of the annotation process we have decided to analyse the number of annotations that match between the annotators. Matches were computed considering different sizes of the partial match. The size of the partial match is the number of characters that are different between two overlapped annotations with the same category. Taking the beginning and end of two annotations, we compute the size of the overlap between them; in case of a perfect match (0), the beginning and end of the two annotations are equal.

In addition, we have also checked the level of inter-agreement among annotators. Although Cohen's Cappa (Cohen, 1960) is the standard annotation reliability measure for many classification tasks, it has been shown not a relevant measure for token-level annotation tasks, like NER (Grouin et al., 2011; Hripcsak & Rothschild, 2005). Therefore, as suggested by Brandsen et al. (2020), we have used the F score – a measure of a test's

[6]We used *doccano* as the platform for the annotation: https://github.com/doccano/doccano.

accuracy originally formulated for binary classification – for this purpose. In the following, we provide the definitions of the applied measure, that is:

$$\mathsf{F} = 2\frac{\mathsf{Precision} \times \mathsf{Recall}}{\mathsf{Precision} + \mathsf{Recall}}.$$

In particular, the Precision is the number of true positive results divided by the number of all positive results, including those not identified correctly. The Recall is the number of true positive results divided by the number of all samples that should have been identified as positive. In formulas: $\mathsf{Precision} = \frac{\mathsf{TP}}{\mathsf{TP+FP}}$, $\mathsf{Recall} = \frac{\mathsf{TP}}{\mathsf{TP+FN}}$, where TP denotes the number of true positive results, FP denotes the number of false positive results, FN denotes the number of false negative results. In each comparison, the annotations of one annotator are considered the gold standard, while those of the other annotator are considered for evaluation.

Table 1 reports both the number of annotations that match between the annotators (column 2) and the F score (column 3) measured by varying the size of the partial match (column 1). As expected, the higher the size of the partial match, the higher the number of annotations that match between the annotators. However, the number of annotations that match between the annotators achieves an elbow value with the size of the partial match set equal to 5. In addition, we note that we obtained a high F score that is greater than 0.79 also with the exact match. This shows good quality of the annotations.

For the corpus preparation, a third annotator, also a domain expert, solved cases in which the two annotators did not agree (the two annotators chose a different label for the same entity, which also includes the case in which one of the two missed the annotation). Statistics about the final dataset are reported in Table 2. The table reports the number of documents, number of GDPR data breaches, number of NER annotations and number of annotated entities for each category. The most frequent categories are: *administration, organization, health* and *person*.

By considering that Article No. 4 of GDPR defines personal data as any piece of data that is enough to identify an individual, and that Article No. 2 of GDPR declares that data ceases to be personal when they are made anonymous, and an individual is no longer identifiable, we performed a process of anonymization that replaced any identified or identifiable information with artificial identifiers. Notice that Recital 26 of GDPR defines anonymous information, as «information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable». The GDPR does not apply to anonymized information. To achieve the anonymization goal, we adopted a pipeline that includes the following steps:

1. Each person name n_i was replaced by another name n_r taken randomly from a list of the most common Italian surnames and first names. In the same document an occurrence of the name n_i was replaced always by the same n_r.
2. Each personal id number or Italian fiscal code was replaced by a random sequence of numbers or a random fiscal code, respectively.

Table 1 The number of annotations that match between the annotators and the computed F score

partial match size	#matches	F score
0	3,296	0.7981
3	3,707	0.8976
5	3,886	0.8978
∞	3,985	0.9649

Table 2 Corpus statistics

#documents	90
#GDPR data breach	45
#NER annotations	4,188
#AMM	2,094
#ORG	983
#MED	527
#PER	400
#OM	104
#LOC	80

3. Each location entity was searched in a list of cities and regions, and then it was randomly replaced by another city or region. The list of replacements contains only cities with more than 10,000 inhabitants to exclude not common city names.
4. Each legal reference[7] was replaced by a placeholder LEX_u. Also, in this case, for each a_i, the replacement was always the same in the whole corpus.

Finally, all documents were manually checked. This anonymised corpus was intended to serve for evaluating the performance of the proposed INTREPID framework, while the original corpus was discarded.

4 The INTREPID framework

The idea behind the proposed framework is to model the problem of data protection of public documents as a binary text classification task. Each document in our corpus is labelled as compliant or non-compliant with the GDPR standards. Figure 2 shows a sketch of the INTREPID architecture.

Since the corpus was provided in PDF format, the first step involves extracting textual content from the PDF. Then, a cleaning step is applied to the extracted content, in order to remove no valid characters and correct OCR errors. For extracting the text from PDF, we rely on the Apache Tika library;[8] when the text is not directly available in the PDF, we perform the OCR by using the software Tesseract.[9] Moreover, not valid characters are found through regular expressions. In particular, we detect long sequences of non-alphanumerical characters. An NLP pipeline processes the cleaned text. Note that the proposed framework does not rely on a predefined pipeline, but it is possible to integrate any pipeline that can produce the CoNLL format[10]. The CoNLL format is a textual file in which each line represents a single word with a series of tab-separated fields. In particular, each token in the sentence is denoted by an index (first column) corresponding to the token position in the sentence (starting from 1). The other columns are the features extracted by the pipeline, such as the token, the lemma, the PoS-tag and the IOB2 tag[11] of the entity recognizer. In particular,

[7] Legal references were extracted by the Linkoln tool https://gitlab.com/IGSG/LINKOLN/linkoln.
[8] https://tika.apache.org/
[9] https://github.com/tesseract-ocr/tesseract
[10] https://ufal.mff.cuni.cz/conll2009-st/task-description.html
[11] https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)

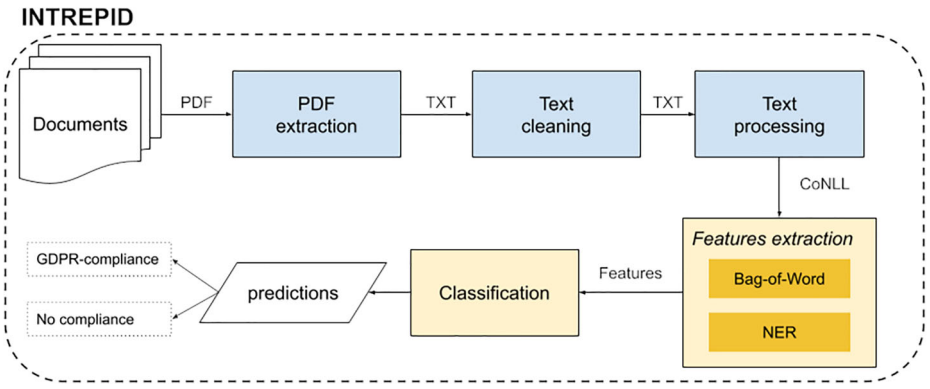


Fig. 2 The INTREPID framework architecture

in our framework, the produced CoNLL format must include the following text operations: tokenization, lemmatization, stop words removal, PoS-tagging and entity recognition.

In the current version of the framework, we provide support for two NLP pipelines: SpaCy[12] and Tint (Palmero Aprosio & Moretti, 2018). The motivation for this choice is that they provide the best support for the Italian language and do not require further configuration or development steps. Moreover, both pipelines can easily perform entity recognition which is the core component for building features based on named entities. For recognising domain-dependent entities (*health, administration and omissis*), we exploit specific gazetteers built by a domain expert. More in detail, the health and administration gazetteers are composed of keywords that identify medical and administrative terms, respectively, while the omissis gazetteer is a list of linguistic expressions identifying omissions in the text. The next step involves extracting features for the classification. In this step, the output of the NLP pipeline in the CoNLL format is used to extract features useful to train the classifier. We exploit two kinds of features: classical Bag-of-Word (BoW) and others based on the output of the NER. We believe that the occurrences of entities can help the classifier to detect non-compliant documents. A detailed description of the feature extraction step is reported in Section 4.1. The extracted features are used to train a classifier, which is then used to predict the class of new documents. The INTREPID framework can be easily adapted to support different NLP pipelines and several tasks involving text categorisation in the PA.

Considering the small size of the corpus and the number of the involved features, in the current implementation of the proposed framework, we provide support for three base classifiers: Linear Support Vector Machines (Joachims, 1998), Random Forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016). These classification algorithms were selected as various studies proved that they commonly outperform several competitor classifiers (Bansal & Kaur, 2018; Ghosh et al., 2021).

To build the best classifier and measure the contribution of each feature group, we perform an evaluation in which a feature group selection step is exploited. In particular, we build all the possible combinations of features and test each one by exploiting the manually annotated corpus. For each combination, we also select the best parameters. Details about the evaluation are reported in Section 5.

[12] https://spacy.io/

4.1 Feature extraction

For the feature extraction stage, we consider both BoW-based features and NER-based features.

4.1.1 BoW-based features

The BoW feature model (Joachims, 1998) is a basic strategy commonly used for feature generation in NLP. To represent each textual document in the BoW model, we first transform the document into a "bag of words". Then, we calculate frequency features from the BoW vocabulary by counting the number of times a term of the vocabulary appears in the document.

In the proposed framework, the BoW vocabulary is populated with the unique words that appear at least once in the corpus. We use lemmatization to convert the vocabulary words into their base form, by also taking into account the part-of-speech context (e.g. the base form of "dichiarato" is "dichiarare" in the Italian text reported in Fig. 1a, while the base form of "declared" is "declare" in the English translation reported in Fig. 1b). In addition, we lowercase words to overcome the uppercase issue (e.g. the proper name "Paolo" is lowercased as "paolo"). Stop-words (i.e. articles, prepositions, pronouns and conjunctions) are removed.

For example, the BoW model of the Italian text in Fig. 1a comprises the following words:

```
paolo
rossi
copenaghen
roma
xxxxxx
sclerosi
multipla
roma
...
```

The BoW feature **`paolo`** occurs once, while the BoW feature **`roma`** occurs twice in the Italian text in Fig. 1a

4.1.2 NER-based features

We consider three groups of NER-based features.

Group 1: Bag of Named Entities (BoNE) The BoNE feature model extends the BoW model, in order to generate NER-based features. To represent each textual document, we first transform it into a "bag of named entities" by using the NER tool to recognise named entities in the text corpus. Then, we calculate frequency features from named entity categories (administration, health, location, omissis, organisation and person) by counting the number of times a named entity category appears in the document.

For example, let us consider the BoNE model of the Italian text in Fig. 1a. It comprises the following bag of named entities:

PERSON	**paolo rossi**
LOCATION	**copenaghen**
LOCATION	**roma**
OMISSIS	**xxxxxxx**
HEALTH	**sclerosi multipla**
OMISSIS	**xxxxxxx**
HEALTH	**fisioterapia**
ORGANIZATION	**ragioneria**
LOCATION	**roma**
PERSON	**maria verde**

The BoNE features are generated by computing the frequency of each named entity category in the previous BoNE model. In the example, we determine that ADMINISTRATION does not appear, HEALTH appears twice, LOCATION appears three times, OMISSIS appears twice, ORGANIZATON appears once and PERSON appears twice. Note that, in the current version of the proposed framework, the BoNE model allows us to extract six features, one for each named entity category, from each document.

Group 2: Bag of Named Entities bi-Grams (BoNNEG) The BoNNEG model is based on the idea of representing each document as a bag of bi-grams of neighbour named entities (i.e. adjacent entities that appear in the same context). To represent each textual document in the BoNNEG model, we first transform the document into a "bag of neighbour named entity bi-grams". Then, we count the frequency features from the named entity bi-grams. Let us consider the named entity bi-gram $\mathbf{X}-\mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are named entity categories recognised by the NER. The bi-gram $\mathbf{X}-\mathbf{Y}$ occurs in a document if and only if there exists an entity x named with $\mathbf{X}$ and an entity y named with $\mathbf{Y}$ that have been identified in the text by the NER so that the distance between x and y is max_d at most, measured as the number of tokens between x and y. The distance is computed by ignoring stop-words. In this case study, max_d is set equal to 5 according to some domain knowledge on the expected distance between relevant entities in PA documents.

For example, let us consider the document reported in Fig. 1a, the BoNNEG model of this text comprises the following bag of neighbour named entity bi-grams:

PERSON-LOCATION	**paolo rossi-copenhaghen**	d=1
PERSON-LOCATION	**paolo rossi-roma**	d=4
LOCATION-LOCATION	**copenhaghen-roma**	d=2
LOCATION-OMISSIS	**copenhaghen-xxxxxxx**	d=4
LOCATION-OMISSIS	**roma-xxxxxxx**	d=1
OMISSIS-HEALTH	**xxxxxxx-fisioterapia**	d=3
OMISSIS-ORGANIZATION	**xxxxxxx-ragioneria**	d=5
HEALTH-ORGANIZATION	**fisioterapia-ragioneria**	d=1
HEALTH-LOCATION	**fisioterapia-roma**	d=3
HEALTH-PERSON	**fisioterapia-maria verde**	d=4
ORGANIZATION-LOCATION	**ragioneria-roma**	d=1
ORGANIZATION-PERSON	**ragioneria-maria verde**	d=2
LOCATION-PERSON	**roma-maria verde**	d=0

The BoNNEG features are generated by computing the frequency of each named entity bi-gram in the previous BoNNEG model. For example, we determine that the bi-gram PERSON-LOCATION occurs twice in the text since the entity "Copenhaghen" named as LOCATION is distant 1 token from the entity "Paolo Rossi" named as PERSON, as well as the entity "Roma" named as LOCATION is distant 4 non stop-word tokens from "Paolo

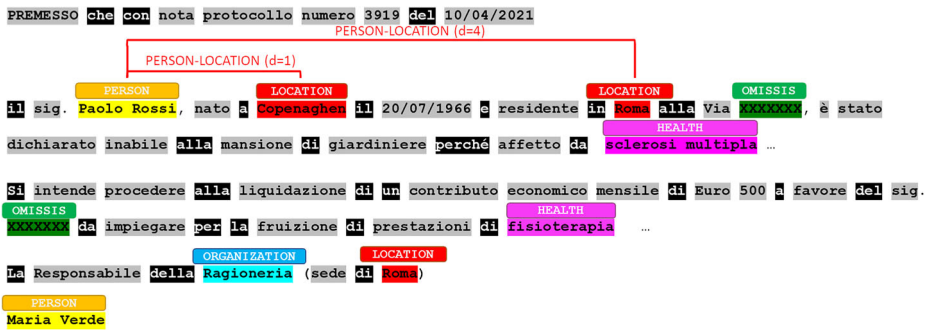


Fig. 3 BoNNEG feature PERSON-LOCATION. Stop-words are highlighted in black

Rossi" (as shown in Fig. 3). Note that the BoNNEG model allows us to extract thirty-six features from each document, one for each bi-gram of named entity categories.

Group 3: Bag of Word per Named Entities (BoWNE) The BoWNE model extracts six features from each document, that is, one feature for each named entity category. To represent each document in the BoWNE model, we combine word information and named entity categorisation. Finally, for each named entity category, we count the number of distinct word tokens named in the considered category that occur at least twice in the text.

For example, the BoWNE model of the document reported in Fig. 1a comprises the following bag of words tagged with a named entity:

paolo rossi/PERSON
copenaghen/LOCATION
roma/LOCATION
xxxxxxx/OMISSIS
sclerosi multipla/HEALTH
xxxxxxx/OMISSIS
fisioterapia/HEALTH
ragioneria/ORGANIZATION
roma/LOCATION
maria verde/PERSON

The BoWNE features are generated by counting duplicated words for each named entity category. Therefore, in the considered example, we determine that LOCATION names only one word (roma) that occurs at least twice in the text, while ADMINISTRATION, HEALTH, OMISSIS, ORGANIZATON and PERSON do not name any word that occurs at least twice in the text. So, the BoWNE features associated with LOCATION and OMISSIS are equal to one, while the BoWNE features associated with ADMINISTRATION, HEALTH, ORGANIZATON and PERSON, respectively, are set equal to zero.

4.2 Implementation details

The framework was implemented in Python and Java. The PDF extraction and text cleaning has been developed in Java using the Apache Tika[13] library for extracting text from PDFs. The text processing module allows the integration of different pipelines for NLP. In

[13] https://tika.apache.org/

Table 3 Parameters and the set of values used during the grid-search

Classifier	Parameters
SVM	C = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
RF	number of estimators = {100, 200, 400, 600, 800}
XGBoost	learning rate = {0.1, 0.2, 0.3}
	max depth = {3, 6, 9}
	number of estimators = {100, 200, 400}

particular, we integrated SpaCy[14] (Python) and Tint (Palmero Aprosio & Moretti, 2018) (Java). The model proposed by SpaCy was trained on the WikiNER corpus (Nothman et al., 2013), while Tint exploits the Italian Content Annotation Bank (I-CAB) (Magnini et al., 2006). Tokenization, stop word removal and lemmatization were performed separately by each NLP pipeline.

The text processing provides the output in the CoNLL format processed by the Feature Extraction module in charge of building both BoW and NER features as described in Section 4.1. Features are the input of the classification module that, together with the evaluation components, have been developed in Python using the scikit-learn tools. For the classification, we adopted: a Linear Support Vector Machine (SVM), a Random Forest (RF) and a XGBoost classifier. Parameters of these classifiers were selected by performing a grid-search with k-fold (k=5). The parameters and the set of values used during the grid-search are reported in Table 3.

5 Experimental evaluation

To evaluate the effectiveness of the proposed framework, we conducted a range of experiments on the text corpus described in Section 3.

5.1 Evaluation goals

The main objective of this experimental study was to explore how INTREPID can be effectively adopted as an AI-based tool to verify the GDPR-compliance of PA documents written in Italian. To this aim, we evaluated how the NER models of both SpaCy and Tint can be used to accurately locate and classify named entities in the considered text corpus. We studied the effect of both BoW-based features and NER-based features on the accuracy of the SVM trained to verify GDPR-compliance of the text corpus. Specifically, we intend to answer the following questions:

Q1: How does the accuracy of the NER stage change by varying the NER model, i.e. SpaCy model and Tint model? (Section 5.3.1)

Q2: How does the use of text pre-processing operations (i.e. lemmatization, stopword removal, lowercase transformation) change the accuracy of the GDPR compliance predictions yielded with BoW-based features? (Section 5.3.2)

Q3: How does each individual NER-based feature group contribute to the accuracy of the GDPR-compliance predictions? (Section 5.3.3)

[14] https://spacy.io/

Q4: Is INTREPID more powerful than its baseline frameworks that separately account for BoW-based features or NER-based features? (Section 5.3.4)

As baseline frameworks we considered BoW and NER, which trained a classifier from BoW-based features and NER-based features, separately. We evaluated the performance of classifiers trained with SVM, RF and XGBoost.

5.2 Evaluation methodology and assessment criteria

We measured the performance of the NER models on the anonymised text corpus by computing the Precision, Recall and F scores of the named entities recognised and classified with both SpaCy and Tint. For both the NER tools, we considered the NER models pre-trained as described in Section 4.2. These three metrics were used to evaluate the predictive ability of a NER tool with respect to a specific entity category i. The higher the F score on the entity category i, the better the balance between Precision and Recall achieved by the NER tool on i. On the contrary, the F score is not so high when one measure is improved at the expense of the other. In addition, we measured the overall performance of NER tools by computing the MacroF, that is,

$$\text{MacroF} = \frac{1}{k}\sum_{i=1}^{k} F_i.$$

This is a standard multi-class metric that is commonly used to evaluate the overall predictive ability of multi-class classification models. It measures the average F score per named entity category i. Note that, in the computation of the MacroF, we give equal weights to each entity category. In this way, we avoided that our evaluation offset the possible impact of imbalanced entity setting.

Similarly, we measured the performance of classifiers trained to verify GDPR compliance of public documents by computing the F scores of document predictions. We evaluated the performance of classifiers trained through the classification algorithms: SVM, RF and XGBoost, by partitioning the anonymized text corpus in training and testing documents according to a leave-one-out cross validation (Hastie et al., 2001). For each trial, we trained the classifier of the compared approaches on the training set (89 folds) and evaluated their ability to predict the GDPR-compliance on the document of the testing set (the hold-out fold). We computed the F score on the GDPR-compliance predictions yielded on all 90 testing trials.

5.3 Results

In this section, we illustrate how the experimental results collected in the evaluation study allow us to address the research questions issued.

5.3.1 NER analysis (Q1)

We start analysing the accuracy performance of both Tint and SpaCy in tagging named entities. We compare the output provided by the two NLP pipelines against the annotation provided by experts in the corpus. In particular, we compare the performance of the two NER tools using two approaches:

– EXACT: the output is correct if both begin and end offsets are equal to the corresponding offset in the annotation;
– OVERLAP: the output is correct if its span overlaps the annotations' span.

Tables 4 and 5 report Precision, Recall and F score, computed for each entity category with EXACT and OVERLAP, respectively. Results computed using EXACT match (Table 4) are very low, especially for SpaCy. These results are expected since the NLP tools were originally trained on a completely different domain. Moreover, the domain taken into account is very specific. On the other hand, we obtain better results if we consider the OVERLAP approach (Table 5). In any case, Tint achieves better results than SpaCy, independently of the approach. Finally, looking at results for each entity category, we can notice that AMM, LOC and ORG are the most difficult categories to recognise, while both tools achieve better performances on PER, MED and OM.

Based on the results of this evaluation, all the subsequent analyses will be conduced by considering Tint as NER tool.

5.3.2 BoW-based classification analysis (Q2)

We proceed exploring the accuracy performance of classifiers trained by accounting for BoW-based features. Classifiers were trained by the classification algorithms: SVM, RF and XGBoost. The tested configurations are denoted as: BoW+SVM, BoW+RF and BoW+XGBoost, respectively . In this experiment, we analysed the effect of lemmatization, stopword removal and lowercase transformation on the accuracy performance of classifiers trained with BoW-based features by varying the classification algorithm.

Results reported in Table 6 show that the highest accuracy performance is achieved in each classification algorithm by applying some text pre-processing operations. However, the adopted text pre-processing operations may have a different effect on the performance of the classification algorithm. Notably, in all classification algorithms, the best accuracy performance is achieved by performing the lowercase conversion of the text. However, in BoW+SVM, the highest accuracy performance is achieved by lower casing the text corpus and removing stopwords. On the other hand, in BoW+RF, the highest accuracy performance is achieved by lower casing the text corpus, performing lemmatization and removing

Table 4 NER accuracy results using EXACT match

Category	Tint			SpaCy		
	Precision	Recall	F	Precision	Recall	F
AMM	0.6196	0.1237	0.2062	0.5892	0.0993	0.1700
LOC	0.2368	0.6750	0.3506	0.0456	0.6000	0.0847
ORG	0.2652	0.2085	0.2335	0.0926	0.0651	0.0765
PER	0.6862	0.6450	0.6649	0.3069	0.3875	0.3425
MED	0.7351	0.6793	0.7061	0.8222	0.6319	0.7146
OM	0.9902	0.9712	0.9806	0.9894	0.8942	0.9394
macroF			0.5237			0.3880

The best MacroF is underlined

Table 5 NER accuracy results using OVERLAP

Category	Tint			SpaCy		
	Precision	Recall	F	Precision	Recall	F
AMM	0.8325	0.1662	0.2771	0.8244	0.1390	0.2378
LOC	0.3026	0.8625	0.4481	0.0598	0.7875	0.1112
ORG	0.5226	0.4110	0.4601	0.2938	0.2065	0.2425
PER	0.8191	0.7700	0.7938	0.4693	0.5925	0.5238
MED	0.8480	0.7837	0.8146	0.9111	0.7002	0.7918
OM	1.0000	0.9808	0.9903	1.0000	0.9038	0.9495
MacroF			0.6307			0.4761

The best MacroF is underlined

Table 6 F score of BoW+SVM, BoW+RF and BoW+XGBoost with respect to pre-processing operations (lemmatization, lowercase transformation and stopword removal)

Conf.	Lemm.	Lowercase	Stopword	F-NC	F-C	MacroF
BoW+SVM				0.7692	0.7692	0.7692
			×	0.8132	0.8132	0.8132
		×		0.7473	0.7473	0.7473
		×	×	0.8352	0.8315	0.8333
	×			0.7473	0.7473	0.7473
	×		×	0.7473	0.7527	0.7500
	×	×		0.7473	0.7473	0.7473
	×	×	×	0.7473	0.7416	0.7444
BoW+RF				0.7363	0.7391	0.7377
			×	0.7363	0.7447	0.7405
		×		0.7253	0.7312	0.7282
		×	×	0.7253	0.7368	0.7311
	×			0.7582	0.7609	0.7596
	×		×	0.7582	0.7755	0.7669
	×	×		0.7692	0.7742	0.7717
	×	×	×	0.7912	0.8000	0.7956
BoW+XGBoost				0.7692	0.7742	0.7717
			×	0.8022	0.8085	0.8054
		×		0.7802	0.7826	0.7814
		×	×	0.7473	0.7473	0.7473
	×			0.7692	0.7692	0.7692
	×		×	0.6923	0.6889	0.6906
	×	×		0.8462	0.8511	0.8486
	×	×	×	0.6264	0.6304	0.6284

NC denotes the label "non-compliant", C denotes the label "compliant". The best results are underlined

stopwords. In BoW+XGBoost, the highest accuracy is achieved by lowercasing the text corpus and performing lemmatization. Finally, we note that the highest accuracy performance with the BoW-based features is achieved by training the XGBoost classifier with the SVM classifier as a runner-up.

5.3.3 NER-based classification analysis (Q3)

Subsequently, we explore the performance of classifiers trained by accounting for NER-based features.

Results reported in Table 7 for NER+SVM, NER+RF and NER+XGBoost show that the accuracy performance achieved using the NER-based features varies with the classification algorithm. In NER+SVM, the highest accuracy performance is achieved in the configuration that trains the SVM classifier by accounting for the bag of named entities of the BoNE feature group. Further accuracy is neither gained by leveraging bi-grams of named entities (BoNNEG feature group) nor combining word information with named entities (BoWNe feature group). On the other hand, in NER+RF, the highest accuracy performance is achieved in the configuration that trains the RF classifier with the word information combined with

Table 7 F score of NER+SVM, NER+RF and NER+XGBoost with respect to NER-based feature groups (BoNE, BoNNEG and BoWNE)

Conf.	BoNE	BoNNEG	BoWNE	F-NC	F-C	MacroF
NER+SVM	×	×	×	0.8352	0.8421	0.8386
	×	×		0.8352	0.8421	0.8386
	×		×	0.8352	0.8421	0.8386
	×			0.8352	0.8421	0.8386
		×	×	0.7473	0.7473	0.7473
		×		0.7473	0.7473	0.7473
			×	0.6593	0.6173	0.6376
NER+RF	×	×	×	0.6703	0.6809	0.6755
	×	×		0.6703	0.6809	0.6755
	×		×	0.7582	0.7660	0.7621
	×			0.7582	0.7660	0.7621
		×	×	0.7582	0.7660	0.7621
		×		0.7582	0.7660	0.7621
			×	0.7912	0.7957	0.7934
NER+XGBoost	×	×	×	0.6374	0.6374	0.6374
	×	×		0.6374	0.6374	0.6374
	×		×	0.7582	0.7609	0.7596
	×			0.7692	0.7789	0.7741
		×	×	0.7692	0.7789	0.7741
		×		0.7692	0.7789	0.7741
			×	0.7692	0.7789	0.7741

NC denotes the label "non-compliant", C denotes the label "compliant". The best results are underlines

the named entities (BoWNe feature group). Finally, in NER+XGBoost, the highest accuracy is achieved by accounting for either the bag of named entities (BoNE feature group) or the bi-grams of named entities (BoNNEG feature group) or the combination of word information with named entities (BoWNe feature group), separately. Notably, the highest accuracy performance with NER-based features is achieved training the SVM classifier having the RF classifier as runner-up.

5.3.4 INTREPID vs baseline analysis (Q4)

Finally, we explore the accuracy performance of the proposed framework, INTREPID, that trained a classifier by accounting for the feature vector obtained by concatenating both BoW-based features and NER-based features, simultaneously. Again we analyse the performance of INTREPID by varying the classification algorithm among SVM, RF and XGBoost. For each classification algorithm, we ran 56 distinct configurations of INTREPID defined by varying the set-up of the text preprocessing pipeline (Lemmatization, Lowercase and Stopword removal) in the BoW-based feature extraction, as well as the selection of the NER-based feature engineering schemes in the NER-based feature extraction. The experimented configurations are described in Table 8.

Figure 4 shows the box-plots of the F score computed for both classes "non-compliant" (F-NC) and "compliant" (F-C), as well as the MacroF metric computed on the configurations of INTREPID run for the classification algorithms: SVM, RF and XGBoost, respectively. These results show that the highest performance is achieved by XGBoost with SVM as runner-up.

We proceed this analysis comparing the accuracy performance of INTREPID to that of the baselines that trained their classifier by accounting for either BoW-based features or NER-based features, separately. Table 9 shows the highest accuracy performance achieved with these three methods. These results confirm that best set-of to be adopted for building both the BoW-based features and the NER-based features changes with the classification algorithm. In addition, they show that NER-based features commonly perform closely to BoW-based features with SVM and RF, while they perform worse than BoW-based features with SVM. In any case, leveraging both BoW-based features and NER-based features within the same classification stage (INTREPID) allows us to learn a classifier able to gain accuracy in verifying GDPR-compliance of PA documents written in Italian. This conclusion can be drawn independently on the classification algorithm, although the highest accuracy performance is achieved with INTREPID when the classifier trained with XGBoost.

Table 8 Configurations of INTREPID

Feature Group	Parameter	Value Range
BoW	Lemmatization	{enabled, disabled}
	Lowercase	{enabled, disabled}
	Stopword	{enabled, disabled}
NER	BoNE	{enabled, disabled}
	BoNNEG	{enabled, disabled}
	BoWNE	{enabled, disabled}

In each configuration, both BoW-based features and NER-based features are computed, so that at least one option among BoNE, BoNNEG and BoWNE is enabled

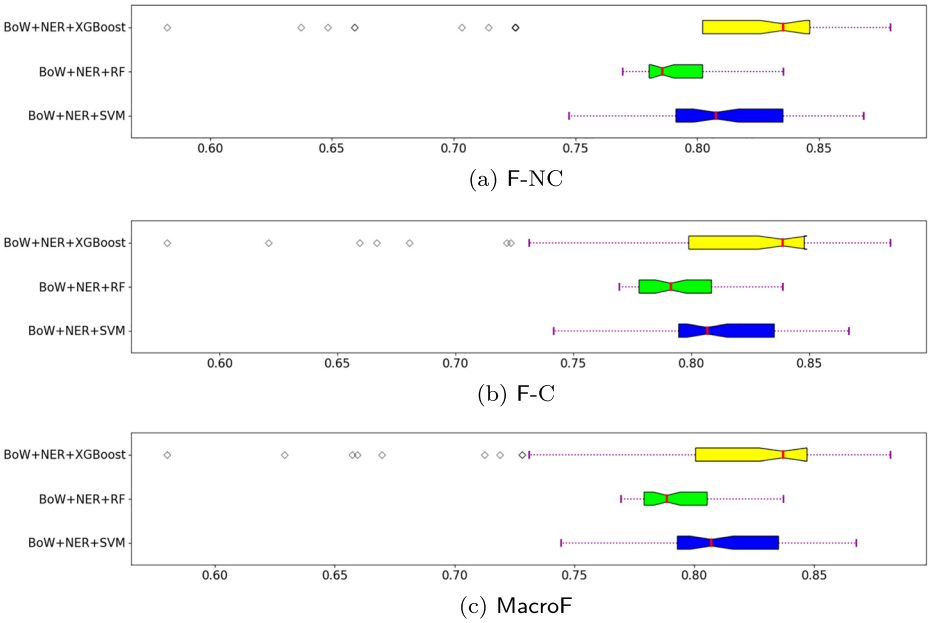


Fig. 4 F-NC (Fig. 4a), F-C (Fig. 4b) and MacroF (Fig. 4c) of INTREPID (Bow+NER) with SVM, RF and XGBoost as classification algorithms

Table 9 MacroF metric of INTREPID vs. BoW-based and NER-based classifiers trained with SVM, RF and XGBoost

Classif.	Method	MacroF	Configuration set-up					
			BoW-based features			NER-based features		
			Le.	Lo.	St.	BoNE	BoNNEG	BoWNE
SVM	INTREPID	<u>0.8673</u>			×		×	
	BoW	0.8333		×	×			
	NER	0.8386				×		
RF	INTREPID	<u>0.8369</u>	×			×		
	BoW	0.7956	×	×	×			
	NER	0.7934						×
XGBoost	INTREPID	<u>0.8816</u>	×	×		×		
	BoW	0.8486	×	×				
	NER	0.7741				×		

"Le" denotes "Lemmatization", "Lo" denotes "Lowercase" and "St" denotes "Stopword". The best results are underlined

6 Discussion

With the goal to support the Italian PA ensuring the GDPR compliance of public documents and securing personal data, we formulated INTREPID, a framework based on AI to automate the detection of breaches of security in PA documents. As a backbone of our framework we used linguistic resources developed for Italian language processing and tuned to the GDPR intelligence. In addition, we defined a text data engineering module based on both Bag-of-Word and NER information, and used machine learning algorithms for classification. Finally, we prepared a corpus of Italian PA documents for both training and evaluation by using an appropriate pipeline to balance the need of replacing any identified or identifiable information with artificial identifiers and the fact that the GDPR check does not apply to anonymized information. An in-depth evaluation performed on the prepared corpus underlined the effectiveness of INTREPID and the set-up of all the components on which it is built on.

Beyond the accuracy performance of results exhibited by INTREPID, several limitations still need to be addressed to make a further step towards the development of an effective tool for reducing the risk of security breaches in PA documents.

Absence of explanation mechanisms. Nowadays the ability to explain decisions of an AI system is of paramount importance, in order to make automatic decision process accepted by the final user. This is coherent with the GDPR assessment of the "right to explanation" of all the decisions, comprising AI-based decisions, which may significantly affect an individual. To this purpose, a future research direction of this study could be devoted to explore eXplainable Artificial Intelligence mechanisms to enrich the data breach alerts with explanations of how the data breach has been discovered in the text.

Locating data breach positions within documents. The proposed framework performs the classification task at document level. It allows us to recognise a PA document that may be non-compliant with the GDPR standards, but this is done without locating the position data breaches within documents.

Data breach variety. The classification model of the proposed framework has been trained with data breaches related to unlawful disclosure of health information. A future research direction could be devoted to generalise the classification model to various data beach categories.

Multi-language support. The proposed framework has been designed for Italian PA documents. However, new multilingual models have recently appeared and have proved to be very accurate in various text classification tasks (Conneau et al., 2020). This may be explored in a multilingual system for data breach detection.

7 Conclusions

In this paper, we present a new AI-based framework to help automating the data protection workflows of the Italian PA. The proposed framework was designed according to the idea that the data protection of public documents can be formulated as a binary text classification problem. Based upon this idea, we prepared a labelled text corpus of public documents that were published online by various municipalities of the Italian PA. The corpus contains text documents that a human expert labelled as GDPR compliant or GDPR non-compliant. We describe an AI framework to learn a text classification model from this labelled text corpus, so that the learned model can be used to predict a new public document as compliant or

non-compliant with the GDPR standards. To this aim, we selected SpaCy and Tint – two NLP tools able to deal with Italian language – and tuned them to the GDPR intelligence. Specifically, we used both the NER tools to process the prepared text corpus and locate named entities of several categories. We introduced three groups of NER features extracted on the occurrence of the recognised named entities. We leveraged these NER features to enrich the traditional BoW representation of the text documents and train a classifier to label documents as compliant or non-compliant with the GDPR standards. We used Linear Support Vector machine, Random Forest and XGboost as classification algorithms.

We assessed the effectiveness of the proposed framework in terms of agreement of the annotation predictions of the NERs with the annotations by domain experts, as well as sensitivity of the accuracy of the text classification model to the groups of extracted features. In particular, the evaluation on the prepared text corpus shows that Tint outperforms SpaCy in this field in terms of agreement of the annotation predictions with the annotations by domain experts. It also shows that the proposed feature extraction stage works reasonably well, since it enables us to train a text classification model that has achieved a valuable accuracy in detecting documents with data breaches with a low number of false positives. This conclusion can be drawn independently of the classification algorithm, although the highest accuracy performance was achieved training a classifier with XGBoost by accounting for BoW-based and NER-based features, simultaneously.

So far, to the best of our knowledge, this study represents the first attempt of combining interdisciplinary competences, in order to develop a framework that may help the Italian PA in the automation (or semi-automation) of the analysis of the GDPR compliance of public documents. The next stage of this research will be to extend the evaluation of the effectiveness proposed framework by including new documents possibly referred to different categories of data breaches in the text corpus and improve on the performance of NER models using our annotated corpus. In addition, there is a need to extend the framework to other categories of personal data, as well as integrate XAI techniques to explain data breach alerts and develop AI techniques to locate data breach positions within documents labelled as non-compliant with the GDPR standards. Finally, we plan to explore the performance of multilingual resources in the GDPR compliance analysis problem.

Acknowledgements We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU, as well as the PON "Governance e capacità istituzionale" 2014–2020 project "Modelli, Sistemi e Competenze per l'implementazione dell'Ufficio per il Processo/Start UPP" (CUP: H29J22000390006), funded by the Italian Ministry for Universities and Research (MIUR).

Author Contributions **Filippo Lorè**: Conceptualization, Methodology, Validation, Data Preparation, Investigation, Writing - original draft, Writing - review & editing. **Pierpaolo Basile**: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing **Annalisa Appice**: Conceptualization, Methodology, Validation, Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration. **Marco de Gemmis**: Conceptualization, Methodology, Validation, Investigation, Writing - original draft, Writing - review & editing, Supervision **Donato Malerba**: Conceptualization, Writing - original draft, Writing - review & editing. **Giovanni Semeraro**: Conceptualization, Writing - original draft, Writing - review & editing

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement.

Code Availability Code that supports the findings of this study and data extracted for training the classification algorithms are available from the corresponding author upon reasonable request.

Declarations

Ethics Approval We declare that this submission follows the policies as outlined in the Guide for Authors. The current research involves no Human Participants and/or Animals.

Competing interests The authors have no financial or proprietary interests in any material discussed in this article.

Conflict of Interests The authors declare that they have no conflict of interest.

References

Adams, A., Aili, E., Aioanei, D., Jonson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Mikmekova, D., Fernandez Valencia, J., & Wechsler, R. (2019). Anonymate: a toolkit for anonymizing unstructured chat data. In *Proceedings of the workshop on NLP and pseudonymisation, pp. 1–7. Finland: Linköping Electronic Press, Turku*.

Al-Abdulkarim, L., Atkinson, K., & Bench-Capon, T. (2016). A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artificial Intelligence and Law*, *24*, 1–49. https://doi.org/10.1007/s10506-016-9178-1.

Attardi, G., Basile, V., Bosco, C., Caselli, T., Dell'Orletta, F., Montemagni, S., Patti, V., Simi, M., & Sprugnoli, R. (2015). State of the art language technologies for italian: the EVALITA 2014 perspective. *Intelligenza Artificiale*, *9*(1), 43–61. https://doi.org/10.3233/IA-150076.

Bansal, A., & Kaur, S. (2018). Extreme gradient boosting based tuning for classification in intrusion detection systems. In M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, & T. Ören (Eds.) *Advances in computing and data sciences, communications in computer and information science, (vol. 905 pp. 372–380)*. Singapore: Springer. https://doi.org/10.1007/978-981-13-1810-8_37.

Biesner, D., Ramamurthy, R., Stenzel, R., Lübbering, M., Hillebrand, L. P., Ladi, A., Pielka, M., Loitz, R., Bauckhage, C., & Sifa, R. (2022). Anonymization of german financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, *13*(2), 151–161. https://doi.org/10.1007/s41060-021-00285-x.

Blume, P. (2016). Impact of the EU general data protection regulation on the public sector. *Journal of Data Protection & Privacy*, *1*(1), 53–63.

Brandsen, A., Verberne, S., Wansleeben, M., & Lambers, K. (2020). Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, pp. 4573–4577. European Language Resources Association (ELRA)*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, T., & Guestrin, C. (2016). Xgboost: a scalable tree boosting system. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.) *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785–794. Association for Computing Machinery (ACM)*. https://doi.org/10.1145/2939672.2939785.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.) *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, pp. 8440–8451. Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.747.

Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H. W., Palka, P., Sartor, G., & Torroni, P. (2018). CLAUDETTE meets gdpr: automating the evaluation of privacy policies using artificial intelligence. SSRN Electronic Journal, 1–59.

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and open problems of legal document anonymization. *Symmetry, 13*(8).

Dadgostari, F., Guim, M., Beling, P. A., Livermore, M. A., & Rockmore, D. N. (2020). Modeling law search as prediction. *Artificial Intelligence and Law*, *29*, 3–34. https://doi.org/10.1007/s10506-020-09261-5.

Datta, P. (2020). Digital transformation of the italian public administration: a case study. Communications of the Association for Information Systems pp. 252–272. https://doi.org/10.17705/1CAIS.04611.

Davari, M., & Bertino, E. (2019). Access control model extensions to support data privacy protection based on GDPR. In C. Baru, J. Huan, L. Khan, X. Hu, R. Ak, Y. Tian, R. S. Barga, C. Zaniolo, K. Lee, & Y. F. Ye (Eds.) *Proceedings of the 2019 IEEE international conference on big data, big data 2019, pp. 4017–4024. IEEE*. https://doi.org/10.1109/BigData47090.2019.9006455.

De Felice, I., Dell'Orletta, F., Venturi, G., Lenci, A., & Montemagni, S. (2018). Italian in the trenches: linguistic annotation and analysis of texts of the great war. In E. Cabrio, A. Mazzei, & F. Tamburini (Eds.) *Proceedings of the 5th italian conference on computational linguistics, CLiC-it 2018, CEUR Workshop Proceedings, (vol. 2253 pp. 1–5)*.

De Martino, G., Pio, G., & Ceci, M. (2022). PRILJ: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments. *Artificial Intelligence and Law*, *30*, 359–390. https://doi.org/10.1007/s10506-021-09297-1.

Di Cerbo, F., & Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. In A. Benczúr, B. Thalheim, T. Horváth, S. Chiusano, T. Cerquitelli, C. Sidló, & P. Z. Revesz (Eds.) *New trends in databases and information systems, ADBIS 2018, communications in computer and information science, pp. 118–126*. Cham: Springer. https://doi.org/10.1007/978-3-030-00063-9_13.

Di Nicola, P., Grossi, P., & Preti, A. (2016). Rethinking the organization of public administration through the enhancement of human resources. The Istat case. RIEDS-Rivista Italiana di Economia, Demografia e Statistica- The Italian Journal of Economic. *Demographic and Statistical Studies*, *70*(1), 17–28.

Dias, M., Bone, J., Ferreira, J., Ribeiro, R., & Maia, R. (2020). Named entity recognition for sensitive data discovery in portuguese. *Applied Sciences*, *10*, 2303. https://doi.org/10.3390/app10072303.

Francopoulo, G., & Schaub, L. P. (2020). Anonymization for the GDPR in the context of citizen and customer relationship management and NLP. In *Proceedings of the of the workshop on legal and ethical issues (Legal2020), pp. 9–14. European Language Resources Association (ELRA)*.

Ghosh, M., Raihan, M. M., Raihan, M., Akter, L., Bairagi, A., Alshamrani, S., & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict liver disease. *Intelligent Automation and Soft Computing*, *29*, 917–928. https://doi.org/10.32604/iasc.2021.017989.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., & Quintard, L. (2011). Proposal for an extension of traditional named entitites: from guidelines to evaluation, an overview. In *Proceedings of the 5th linguistics annotation workshop (The LAW V), pp. 92–100*. USA: Association for Computational Linguistics, Portland, Oregon.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: automated analysis and presentation of privacy policies using deep learning. In *Proceedings of the 27th USENIX conference on security symposium, SEC'18* (pp. 531–548). USA: USENIX Association.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning. Springer Series in Statistics*. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hoofnagle, C. J., van der Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, *28*(1), 65–98. https://doi.org/10.1080/13600834.2019.1573501.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298. https://doi.org/10.1197/jamia.M1733.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec, & C. Rouveirol (Eds.) *Proceedings of 10th european conference on machine learning: ECML-98, lecture notes in computer science, (vol. 1398 pp. 137–142). Berlin, Heidelberg: Springer*. https://doi.org/10.1007/BFb0026683.

Kingston, J. (2017). Using artificial intelligence to support compliance with the general data protection regulation. *Artificial Intelligence and Law*, *25*, 429–443. https://doi.org/10.1007/s10506-017-9206-9.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., & Sprugnoli, R. (2006). I-CAB: the italian content annotation bank. In *Proceedings of the 5th international conference*

on language resources and evaluation (LREC '06), pp. 963–968. Italy: European Language Resources Association (ELRA), Genoa.

Mc Cullagh, K., Tambou, O., & Bourton, S. (eds.) (2019). *National adaptations of the GDPR*, 1st edn. Blogdroiteuropéen: Collection Open Access Book.

Meszaros, J., & Ho, C. (2021). AI research and data protection: can the same rules apply for commercial and academic research under the GDPR? *Computer Law & Security Review*, *105532*, 41. https://doi.org/10.1016/j.clsr.2021.105532.

Mozes, M., & Kleinberg, B. (2021). No intruder, no validity : evaluation criteria for privacy-preserving text anonymization . Preprint at arXiv:2103.09263.

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, *194*, 151–175. https://doi.org/10.1016/j.artint.2012.03.006.

Palmero Aprosio, A., & Moretti, G. (2018). Tint 2.0: an all-inclusive suite for NLP in italian. In *Proceedings of the 5th italian conference on computational linguistics, CLiC-it 2018, CEUR workshop proceedings, (vol. 2253, pp. 1–7)*.

Passaro, L. C., Lenci, A., & Gabbolini, A. (2017). Informed PA: a NER for the italian public administration domain. In R. Basili, M. Nissim, & G. Satta (Eds.) *Proceedings of the 4th italian conference on computational linguistics, CLiC-it 2017, CEUR Workshop Proceedings, Vol. 2006*.

Ricci, A. (2018). E-government, transparency and personal data protection.: a new analysis' approach to an old juridical issue. *Central and Eastern European eDem and eGov Days*, *325*, 125–135. https://doi.org/10.24989/ocg.v325.11.

Romano, M. F., Baldassarini, A., & Pavone, P. (2020). Text mining of public administration documents: preliminary results on judgments. In D. F. Iezzi, D. Mayaffre, & M. Misuraca (Eds.) *Text analytics: advances and challenges. proceedings of the 14th international conference on the statistical analysis of textual data (JADT 2018), studies in classification, data analysis, and knowledge organization, pp. 117–126. Cham: Springer*. https://doi.org/10.1007/978-3-030-52680-1_10.

Sartor, G., & Lagioia, F. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. European Parliamentary Research Service. https://doi.org/10.2861/293.

Savic, D., & Veinovic, M. (2018). Challenges of general data protection regulation (GDPR). In *Proceeding of the 5th international scientific conference on information technology and data related research, sinteza 2018, pp. 23–30. Serbia: Singidunum University, Belgrade*. https://doi.org/10.15308/Sinteza-2018-23-30.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022.

Silva, P., Gonçalves, C., Godinho, C., Antunes, N., & Curado, M. (2020). Using natural language processing to detect privacy violations in online contracts. In *Proceedings of the 35th annual ACM symposium on applied computing, SAC 2020, pp. 1305–1307. New York: Association for Computing Machinery (ACM)*, 10.1145/3341105.3375774.

Sovrano, F., Vitali, F., & Palmirani, M. (2020). Modelling GDPR-compliant explanations for trustworthy ai. In A. Kò, E. Francesconi, G. Kotsis, A. M. Tjoa, & I. Khalil (Eds.) *Electronic Government and the Information Systems Perspective. Proceedings of the 9th international conference on electronic government and the information systems perspective, EGOVIS 2020, lecture notes in computer science, (vol. 12394 pp. 219–233). Cham: Springer*. https://doi.org/10.1007/978-3-030-58957-8_16.

Stamova, I., & Draganov, M. (2020). Artificial intelligence in the digital age. In *Proceedings of the international scientific conference "digital transformation on manufacturing, infrastructure and service", IOP conference series: materials science and engineering, vol. 940*. https://doi.org/10.1088/1757-899X/940/1/012067.

Sánchez, D., Viejo, A., & Batet, M. (2021). Automatic assessment of privacy policies under the GDPR. Applied Sciences 11(4). https://doi.org/10.3390/app11041762.

Tagarelli, A., & Simeri, A. (2021). Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code. *Artificial Intelligence and Law*, *30*, 417–473. https://doi.org/10.1007/s10506-021-09301-8.

van der Aalst, W. M. P. (2016). *Process Mining- Data Science in Action*, 2nd edn. Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-662-49851-4.

van Engers, T. M. (2005). Legal engineering: a structural approach to improving legal quality. In A. Macintosh, R. Ellis, & T. Allen (Eds.) *Proceedings of the 25th SGAI international conference on innovative techniques and applications of artificial intelligence, AI-2005* (pp. 3–10). London: Springer. https://doi.org/10.1007/1-84628-224-1_1.

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. Preprint at arxiv:1910.11470.

Zaman, R., Cuzzocrea, A., & Hassani, M. (2019). An innovative online process mining framework for supporting incremental GDPR compliance of business processes. In C. Baru, J. Huan, L. Khan, X. Hu, R. Ak, Y. Tian, R. S. Barga, C. Zaniolo, K. Lee, & Y. F. Ye (Eds.) *Proceedings of the 2019 IEEE international conference on big data, big data 2019, pp. 2982–2991*. https://doi.org/10.1109/BigData47090.2019.9005705.

Zaman, R., & Hassani, M. (2020). On enabling GDPR compliance in business processes through data-driven solutions. *SN Computer Science*, *1*(4), 210. https://doi.org/10.1007/s42979-020-00215-x.

Affiliations

Filippo Lorè[1] · Pierpaolo Basile[1] · Annalisa Appice[1,2] · Marco de Gemmis[1] · Donato Malerba[1,2] · Giovanni Semeraro[1]

Filippo Lorè
filippo.lore@uniba.it

Pierpaolo Basile
pierpaolo.basile@uniba.it

Marco de Gemmis
marco.degemmis@uniba.it

Donato Malerba
donato.malerba@uniba.it

Giovanni Semeraro
giovanni.semeraro@uniba.it

[1] Department of Informatics, Università degli Studi di Bari Aldo Moro, via Orabona, 4 - 70125, Bari, Italy

[2] Consorzio Interuniversitario Nazionale per l'Informatica- CINI , via Orabona, 4 - 70125, Bari, Italy