IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 December 2022, accepted 11 January 2023, date of publication 30 January 2023, date of current version 6 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240089

RESEARCH ARTICLE

Is Your Model Sensitive? SPeDaC: A New Resource for the Automatic Classification of Sensitive Personal Data

GAIA GAMBARELLI[1,2], ALDO GANGEMI[1,3], AND ROCCO TRIPODI[4]

[1]FICLIT, University of Bologna, 40125 Bologna, Italy
[2]Ellysse srl., 42124 Reggio Emilia, Italy
[3]ISTC-CNR, 00185 Rome, Italy
[4]LILEC, University of Bologna, 40125 Bologna, Italy

Corresponding author: Gaia Gambarelli (gaia.gambarelli2@unibo.it)

This work was supported in part by University of Bologna, in part by Ellysse srl., and in part by TAILOR Project (Foundations of Trustworthy AI—Integrating Reasoning, Learning and Optimization), H2020-ICT-2019-3, EC GA952215.

ABSTRACT In recent years, there has been an exponential growth of applications, including dialogue systems, that handle sensitive personal information. This has brought to light the extremely important issue of personal data protection in virtual environments. Sensitive information detection (SID) covers different domains and languages in literature. However, if we refer to the personal data domain, the absence of a shared standard benchmark makes comparison with the state-of-the-art difficult for this task. To fill this gap, we introduce and release SPeDaC, a new annotated resource for the identification of sensitive personal data categories in the English language. SPeDaC enables the evaluation of computational models for three different SID subtasks with increasing levels of complexity. SPeDaC 1 regards binary classification, a model has to detect if a sentence contains sensitive information or not; in SPeDaC 2 we collected labeled sentences using 5 categories that relate to macro-domains of personal information; in SPeDaC 3, the labeling is fine-grained and includes 61 personal data categories. We conduct an extensive evaluation of the resource using different state-of-the-art-classifiers. The results show that SPeDaC is challenging, particularly with regard to fine-grained classification. Classifiers based on the transformer architectures achieve good results on SPeDaC 1 and 2 but have difficulties to discern among fine-grained classes in SPeDaC 3.

INDEX TERMS Personal data classification, privacy protection, sensitive data corpus, sensitive information detection, sensitive personal data, transformer models.

I. INTRODUCTION

In recent years, there has been an exponential growth of applications, including dialogue systems that handle sensitive personal information [1], [2], [3]. Identifiable individuals can explicitly or implicitly reveal inferable personal information from the texts they write and from the information they share daily online (in blogs, public pages, social media, etc.). The context in which personal information can be expressed concerns not only public online environments but also private interactions, in which, sometimes, the sharing of such information is deemed necessary. Exchanges of emails in company structures, virtual interactions between users and operators of customer service, or even the use of applications based on human-robot (H-R) interactions are all scenarios in which the management of personal information is important. In online conversations and unstructured text, for example, the loss of privacy can be very high and the average cost of data breaches has increased over the years [4]. The loss of personal information to 3rd parties can have both legal and economic repercussions on the users and managers of the service, and, in social terms, on the individuals directly involved. Finally, it is estimated that 80% of the data currently disseminated on the Internet is of an unstructured type [5] i.e., data not present in a relational database, which can be presented in an irregular and contextual form.

According to General Data Protection Regulation (GDPR, 2018) [6], the right to privacy regarding sensitive personal

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

data is claimed; managing privacy and understanding the processing of personal data has become a fundamental right, especially within the European Union (GDPR, Recital 6) [7]. Following the regulatory definition, 'personal data' means 'any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person' (GDPR, 4.1) [8].

Consequently, many studies have focused on protecting privacy in virtual spaces from several points of view e.g., data sanitization and anonymization methods. Models of data sanitization by using deletion operations on transactions are one of the most common approaches in privacy preserving data mining (PPDM) [9]. PPDM techniques 'allow the extraction of information from data sets while preventing the disclosure of data subjects' identities or sensitive information' [10]. Data sanitization generally aims to hide sensitive information applying minimal side effects and keeping the original database as authentic as possible [11]. Several methods are applied to input data e.g., data perturbation [12], [13], cryptography [14], [15], and anonymity with different techniques methods [16], [17], [18], [19]. A recent PPDM study [20] introduces PACO2DT, an ACO-based multiobjective model which uses transaction deletion to hide sensitive and confidential information. The information to hide is defined by an expert in the industry in form of an input-sensitive itemset, deleted and distributed to the IoT devices for configuration. In a previous study [21], where the hiding-missing-artificial utility (HMAU) algorithm is introduced to address the PPDM problem, the authors propose in future work to extend the sensitive itemsets to be hidden to the sensitive association rules and to decrease the confidence of these rules. A type of extension is proposed introducing high-utility itemset mining (HUIM), a model which discovers itemsets reporting a high profit in transaction databases [22]. The authors introduce, as an extension of PPDM, the preserving utility mining (PPUM) that hides sensitive high-utility itemsets (SHUIs) considering their profit. PPDM and PPUM algorithms are included in the proposed interface privacy-preserving and security mining framework (PPSF) [23]. One of the greatest risks of failure of these models is the loss of information. Even before the failure of the anonymization algorithm, there may be a missing identification of sensitive information.

Sensitive information detection (SID) is a subpart of data leak detection (DLD) that deals with the automatic identification of sensitive information. The work also contributes to improving data loss prevention (DLP) systems and industrial problems designed to help businesses to avoid data breaches, presenting a way to train, classify and perform the classification of sensitive text [24]. Most of the current tools offer DLP services for the automatic identification of personally identifiable information (PII) [25], [26], [27]. This paper addresses the challenge of identifying complex personal information in unstructured text. Words are sensitive or not sensitive depending on their context. Using different expressions in natural language, the same keyword can acquire a sensitive or non-sensitive character.

Related work in SID is often conducted in different domains or languages; however, frequently, there is no common benchmark or available labeled resources to compare the results with the state-of-the-art methods. We have attempted to fill this gap by introducing and evaluating a new sensitive resource. The datasets are freely available and can be reused for training new models or as a benchmark to compare the results to state-of-the-art models.

At the same time, evaluating the resource, we have a neural networks method based on the transformer architecture [28], which has recently been used in SID tasks and has achieved astonishing performances on standard natural language processing (NLP) tasks. The contributions of this study are as follows:

1) we present SPEDAC (Sensitive Personal Data Categories corpus): a benchmark built and manually labeled for personal data categories (PDCs). The dataset contributes to the detection of sensitive sentences and their classification as PDCs. We implicitly contribute to the evaluation gap [29] and the absence of an available labeled resource concerning personal information;
2) we report the results of several experiments conducted using different state-of-the-art models, including a classifier based on the transformer architectures [30]. We aim to evaluate the SPEDAC dataset, propose a benchmark and analyze the validity of modern neural network approaches to the task of automatic identification of sensitive content.

The article consists of the following sections: Section II is devoted to related work in the automatic identification of sensitive content and the use of transformer neural networks in text classification. In Section III we describe the materials and models involved: in Section III-A we introduce the taxonomy used to define the PDCs; in Section III-B SPEDAC, the constructed and labeled sensitive data corpus, is presented. The resource is evaluated in Section III-C, where we further explore the machine learning models as well as the transformer networks, both used to conduct comparative experiments and validate the efficiency of the latter. In Section IV, we describe the experimental process which includes the feature extraction and the models setup, while in Section V we report the results. Finally, Section VI presents conclusions and future directions of work, and in Section VII you can find an ethical disclosure that concerns the protection from improper use of the resources presented.

II. RELATED WORK

The domain of our study concerns personal data categories. In literature, not so many works focus on this specific domain, considering e.g., basic personal information [31], [32],

personal health information (PHI) [33], ethnic origin, and political opinion information [34].

However, regardless of the type of information considered in the literature, sensitive data can be identified in a rigid and context-less manner or can be disambiguated or inferred from the context. We divided the studies into two macro approaches:

1) **non-context-aware approach**, where sensitive information does not depend on the context in which it appears; for example, a word can be identified as sensitive regardless of the sentence in which it is used;
2) **context-aware approach**, where the sensitivity of the data varies according to the context. Only given the sentence, we can infer the sensitivity of a given word. Assuming the textual unit in which a word appears as context [35], we consider the context of a word to be the sentence in which it appears. Nevertheless, this sensitive context can be extended to paragraphs or documents.

The first non-context-aware approach includes works based on the identification of fixed context with n-gram techniques [24], [36] or rule-based inferences to identify contextless words with sensitivity scores, [37], [38]. The contextualized approach appears in literature with an embedding technique for the recognition of a fixed context [39].

Among the most recent works, we see the use of neural networks; for example, recursive neural networks for automatic paraphrasing applied to the identification of sensitive data [29]. Convolutional neural networks (CNNs) [40] have also been used for the sensitive detection of military and political documents in the Chinese language [41]. Bidirectional Long-Short Term Memory (Bi-LSTM) neural networks [42] have been used in a study conducted in the Chinese language on the unstructured text [43] and for the identification of personal data in an Amharic text [34].

Finally, the field of identification of sensitive data began to take advantage of the transformer architecture [28]. A study conducted in the Spanish language [33] used a BERT-based sequence labeling model to detect and anonymize sensitive data in the clinical domain. Specifically, they used two datasets of medical reports and ran comparison experiments using conditional random field (CRF) and BERT. In the first dataset, the pre-trained BERT model outperformed the other systems, whereas, in the second, it fell at 0.3 F1-score points behind the shared task-winning system, but the authors did not try more sophisticated fine-tuning strategies. A recent study on the English language [32] proposed ExSense, a model named BERT-BiLSTM-attention for extracting sensitive information from unstructured text. The experimental process was conducted on the Pastebin dataset [44], manually labeled with personal information. Personal information refers to identifiable persons, such as name, address, date of birth, social security numbers (SSN), and telephone numbers. This model had an F1 score of 99.15%. As the authors stated, ExSense can identify limited types of sensitive information. The identified categories are, therefore, attributable to very specific entities, often presenting a fixed structure.

A novel framework, Just Fine-tune Twice (JFT), recently proposed [45], aims first to redact in-domain data of the sensitive task and fine-tune the model; second, to privately fine-tune the model on the original sensitive data. The first step allows the model to directly learn information from the in-domain data and to work with a limited amount of data. The goal of the paper is to show the potential outcomes of JFT, and basic sensitive information is treated.

Recent literature on this task highlights the great potential of transformer-based models. However, the type of personal data investigated is often not very challenging. Transformer-based models have never been tested in the English language on such a broad domain of PDCs, as the one presented in this work. Considering the definition of 'personal data' given by the GDPR (Section I), many types of data can be identified textually. Categories such as names, addresses, and telephone numbers can be identified directly, through entities, while there are personal categories, such as health status, preferences, and social status, that can be more complex to identify or infer. Their common feature is that they can be directly or indirectly related to an identifiable person. In addition to investigating the accuracy of PDCs identification, we measured the accuracy that a transformer network can achieve in discriminating between sentences with and without sensitive content, based on the same potentially sensitive linguistic patterns occurring in different sentences that confer sensitivity or not.

Nevertheless, how can we identify the types of sensitive data categories to consider? The World Wide Web Consortium (W3C) [46] created the Data Privacy Vocabulary (DPV) in 2019 [47], a resource aimed at ensuring the interoperability of data privacy, which therefore represents a highly valid reference taxonomy [48]. We have used this as an authoritative reference to identify the categories of personal data (PDCs) to be analyzed. An extension of DPV regarding extended personal data concepts was recently released [49]. The resource will be discussed in more detail in Section III-A. Regarding the second problem, our approach aims to be context-aware; the analysis is therefore a level-sentence, as we will describe in Section III-B.

As mentioned above, one of the major obstacles is the corpora and resources currently available to form and compare sensitive detection models [29]. Some public corpora, that contain sensitive data and which have been used in the sensitive detection literature, are as follows:

1) the Enron email dataset [50], which collects more than 600,000 e-mails from the American Enron Corporation, with approximately 2,720 documents manually labeled by human annotators, lawyers, and professionals in 2010. However, annotations only cover specific topics, such as business transactions, forecasts and projects, actions, and intentions. This dataset was used as an evaluation dataset in related studies [24], [29], [37].

2) The Monsanto dataset [51], published in 2017, consists of secret legal acts. This resource was similarly used for evaluation [29].
3) PII dataset from Pastebin [32]. The authors collected documents from Pastebin, obtaining 144,967 text sequences as training data and identifying 4 types of PII information in text using regular expressions for content-based sensitive information and a BERT-BiLSTM attention model to automatically extract context-based sensitive information from the preprocessed text.
4) Wikipedia dataset. Wikipedia articles or pages are very easy to acquire and contain different types of sensitive information. In a privacy-ensuring study [24] the authors have created a Wikipedia test corpus, randomly sampling 10K Wikipedia articles. In another related work [52] the aim was to establish a framework for measuring the disclosure risk caused by semantically related terms; the authors used Wikipedia pages of individuals e.g., movie stars. They used a manual annotation for sentences on Wikipedia pages relating to PII typically defined by keywords e.g., HIV (state of health), Catholicism (religion), and Homosexuality (sexual orientation).

The Enron corpus could be representative of organizational email conversations, including informal emails between colleagues. However, since it dates back to 2002, it cannot be considered very representative of today's communication style. Although more recent, the Monsanto dataset is a domain-specific corpus that would barely cover many PDCs, other than those closely related to the legal domain. For these reasons, they could not represent a point of reference for the specific identification of personal data. The dataset from Pastebin is not currently available; furthermore, the investigated categories refer to PII, frequently detected through regular expressions or very narrow linguistic patterns. Even the Wikipedia dataset is not publicly available and, in any case, complex sensitive categories are not considered.

This brief survey highlights the clear lack of an available released labeled resource for the task of automatic identification of sensitive personal data. For this reason, this study aims to offer an evaluation and reusable resource as a contribution.

III. MATERIALS AND MODELS

In this section, we deepen the taxonomy used as a reference for the identification of the PDCs analyzed (Section III-A), describe SPEDAC, the corpus built and evaluated (Section III-B), and introduce the machine learning and transformer network models used to conduct the classification experiments (Section III-C).

A. DATA PRIVACY VOCABULARY (DPV)

As introduced in Section II, we decided to pay attention to an authoritative resource, the so-called DPV. This resource enables the expression of machine-readable metadata regarding the use and processing of personal data. It provides terms and definitions according to the GDPR and it is divided into classes and properties. The *basic ontology* describes the first-level classes that define a legal policy for the processing of personal data (see Fig.1).

Following the descriptions given in the latest published version of the resource, we are particularly interested in *Personal Data*, for example, data directly or indirectly associated with or related to an individual. DPV provides the concept of *Personal Data*, and the relation *has Personal Data* to indicate what categories or instances of personal data are being processed. In particular, *Sensitive Personal Data* is a class for indicating personal data that is considered sensitive in terms of privacy and/or impact, and therefore requires additional considerations and/or protection. The Data Privacy Vocabulary-Personal Data (DPV-PD) extension provides an extended DPV personal data taxonomy, where concepts are structured in a top-down schema based on an opinionated structure contributed by R. Jason Cronk from EnterPrivacy [49].

The DPV-PD presents 206 Personal Data Categories (PDCs), according to its most recent release (December 05, 2022). Each category is described by a definition and additional information, such as an IRI (Internationalized Resource Identifier), a source, and its hierarchical relations.

However, not all categories of the DPV can be explored in the same way. A detailed analysis of the resources led us to identify a narrower set of PDCs to be explored through textual analysis. We divided these categories into 5 different types based on their nature and characteristics that can affect their automatic identification. The subdivision is summarized in Table 1.

Macro-categories: The subdivision is based on the ontological organization provided by the DPV [49]. The macro-categories correspond to the high-level categories to which all the more specific PDCs belong. Therefore, their identification was implicit in the identification of nested categories. There are six relevant macro-categories:

1) *Historical*: information about historical data related to or relevant to history or past events e.g., *Life History*.
2) *Financial*: information about finance including monetary characteristics and transactions e.g., *Transactional*, *Ownership*, *Financial Account*.
3) *Tracking*: information used to track an individual or group e.g. location or email e.g., *Location*, *Device Based*, *Contact*.
4) *Social*: information about social aspects such as family, public life, or professional networks e.g., *Family*, *Friends*, *Public Life*.
5) *External* (visible to others): information about external characteristics that can be observed e.g., *Behavioral*, *Physical Trait*, *Physical Characteristic*.
6) *Internal* (within the person): information about internal characteristics that cannot be seen or observed e.g., *Preference* and *Knowledge Beliefs*.

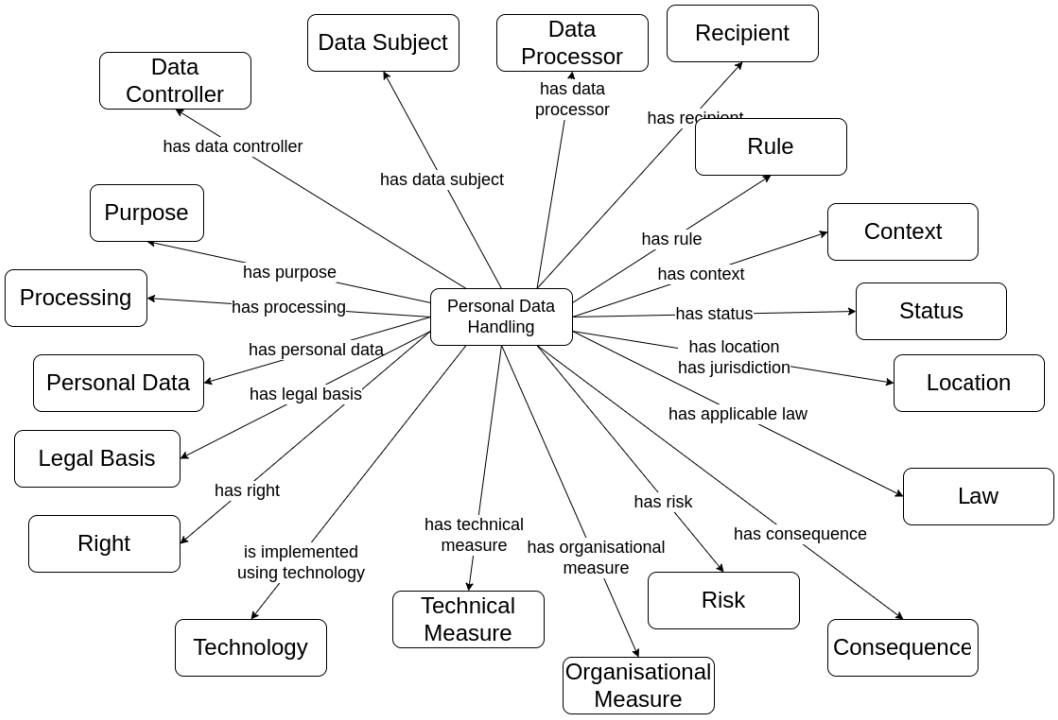


FIGURE 1. Base Vocabulary DPV. The figure can be found in the DPV documentation [49] Copyright © 2022 the Contributors to the Data Privacy Vocabulary (DPV) Specification, published by the Data Privacy Vocabularies and Controls Community Group under the W3C Community Contributor License Agreement (CLA).

Special Category Personal Data, cited as a subtype of *Sensitive Personal Data*, is added. The macro-category is based on GDPR Article 9, even if it considers all sensitive special categories whose use is prohibited or regulated with an additional legal basis for justification. Some PDCs include *Health*, *Mental Health*, and *Disability*.

Recently, *Household* and *Profile* have also been identified as macro-categories but do not present nested categories.

Categories identifiable through textual analysis: These are categories that can be frequently expressed through text and whose expressions can be syntactically complex. They are not alphanumeric sequences or codes that are easily identifiable through regular expressions but can be expressed in natural language depending strongly on the combination of words. Let's take, for example, the *Age* category, whose definition is: '*information about an individual's age*'. Information about an individual's age can be expressed in *n* different ways, such as: '*I'm 17 years old*' or '*I was born in 2005*' or '*In 2010 I was only 5 five years old*': textual elements are crucial for its identification. We first investigated these categories.

Broad-boundary categories: These categories can be defined as characterized by (i) a high degree of vagueness, (ii) a high degree of extension and applicability, and (iii) whose sensitivity classification is characterized by a high degree of ambiguity. For example, *Intention* is a sub-category of *:Preference:Internal* and refers to information about an individual's intentions. These categories, owing to their conceptual complexity, have not been treated as priorities. However, reflections on the future developments of the work are reserved for them.

Uniquely identifiable categories: These are categories easily identifiable through regular expressions and fixed sequences, e.g., *Credit Card Number*, *Tax Code*. This type of category (PII) has been heavily extensively in the literature. Tool markets offered by large companies, e.g., Microsoft [53] can already be found. It seemed appropriate to focus our analysis on the most challenging and least explored categories, which could at the same time give us the possibility of analyzing more complex and context-aware identification techniques.

Categories identifiable mainly through non-textual elements: These depend completely or largely on non-textual elements and it is therefore difficult, if not impossible, to identify them in this sense. An example may concern the *Fingerprint* category: '*information related to an individual's fingerprint used for biometric purposes*'.

Considering the categories identifiable through textual elements, most of the PDCs belong to *Special Data*, *Social* and *External* macro-categories. In general, the ontological structure arrives at four levels of hierarchy. Some categories, in the analysis and consequent construction of the corpus, were merged by similarity, e.g., *Physical Characteristic* and *Physical Trait*, or because they are not strictly necessary specifications of a more generic category, e.g., *Family* and *Family Structure*. A list of identified PDCs labels is provided in Table 5.

TABLE 1. Analysis of the 206 PDCs of the DPV.

N.	Type	Examples
6	Macro-categories	FINANCIAL, EXTERNAL, INTERNAL, HISTORICAL, SOCIAL, TRACKING
90	Identifiable through textual analysis	AGE, FAVORITE, HEALTH
26	Broad boundaries categories identifiable through textual analysis	ATTITUDE, INTEREST, INTENTION
30	Uniquely identifiable	BANKACCOUNT, BLOODTYPE, CREDITCARD
54	Identifiable mainly through non-textual elements	BIOMETRIC, CALLLOG, FINGERPRINT

B. SPeDaC: A SENSITIVE DATA CORPUS

Given the lack of publicly available datasets for sensitive data identification, the first aim of our study is to fill this gap and develop a labeled resource for the task.

Personal data in informal online conversations are the domain context of interest. The TenTen corpus family is a large resource, composed of texts collected from the World Wide Web [54]. TenTen corpora are available in more than 40 languages. The most recent version of the English TenTen corpus (enTenTen2020) consists of 36 billion words. The texts were downloaded from 2019 to 2021. The sample texts were manually checked and content with poor-quality text and spam was removed. These come from different web domains (the UK domain .uk, Australian domain .au, Canadian domain .ca, US domain .us, New Zealand domain .nz, and the EU domain .eu) and different textual genres (news, discussion, blog, legal) and topics (reference, society, arts, technology, business, sports, science, health, home, recreation, games); 6.8% of the corpus comes from English Wikipedia pages.

For our experiments, we created two different corpora, that were manually labeled by the authors. Both corpora present a sentence-level annotation using INCEpTION as an annotation platform [55] and the WebAnno TSV v3.3 annotation format. The datasets are available on a GitHub repository and are shared subject to a declaration of the purposes of use (see Section VII): https://github.com/Gaia-G/SPeDaC-corpora.

1) SPeDaC 1

Identification and discrimination of sensitive sentences from non-sensitive sentences. The dataset counts 10,675 sentences and has two target labels:

1) **0 (NON-SENSITIVE)** to indicate sentences without sensitive content.
2) **1 (SENSITIVE)** to indicate sentences with sensitive content.

For each fine-grained category (see Table 5), we collected sensitive and non-sensitive examples in a balanced manner i.e., considering approximately the same number of examples for each of the two classes. Non-sensitive examples correspond to sentences that contain the same linguistic patterns found in sensitive sentences but in a context that does not confer sensitivity. We can distinguish between two types of linguistic constraints chosen as selection criteria: (i) general constraints and (ii) specific constraints for every PDC. General constraints take into account the importance of the relationship between a PDC and the subject to which it refers. We assume that the identifiable subject (e.g., account, the device used) often corresponds to the person who writes ('I'). The specific linguistic constraints concern multi-word expressions that could better represent every PDC, e.g., the construction '*[have] ... years old*' which recurs to represent the AGE category. As shown in the examples in Table 2, specific constraints are present in sensitive and non-sensitive sentences, whereas the cited general constraint which refers to a first-person subject characterizes only sensitive sentences. Thus, adversarial sentences can better represent ambiguous cases, in which PDCs are present in a non-sensitive context. The adversarial constraints representing the aforementioned cases can consist of citation expressions, e.g., '*[he] [say]*', '*[article][say]*', '*[he][state]*' etc., or expressions concerning the dimension of unreality or supposition, e.g., verbs as '*suppose*', '*imagine*', '*guess*', '*hope*', or adverbs as '*maybe*', or related to the joke e.g., '*just kidding*', '*I [be] joking*'.

2) SPeDaC 2

Identification of the PDC macro-categories within sensitive sentences. The 5,133 sentences in the dataset represent the fine-grained PDCs considered in a balanced manner i.e., approximately the same number of examples for each fine-grained category has been taken into account. Specifically, the aim was to collect 100 sentences from each fine-grained PDC. For some PDCs the retrieval of 100 sensitive sentences was difficult and they are therefore less represented in the corpus, e.g., *Criminal*, *Criminal Conviction*, *Criminal Charge*, *Disciplinary Action*, *Income Bracket*, *Privacy Preference*, *Professional Evaluation*, *Professional Interview*, *Salary*, *Skin Tone*. For every sentence its macro-category has been retrieved, presenting total 5 different labels, which are the following:

1) Special Category Data
2) Financial and Tracking
3) Social
4) Internal
5) External

The category Historical has been excluded because of its inconsistency (it is a superclass only of *Life History* PDC, which is a broad-boundary category).

The percentage of representation of the macro categories in the corpus, which depends on the number of specific categories included, is presented in Table 5.

TABLE 2. Examples from SPEDAC 1.

Sentence	Label
hey! I'm 33 years old now.	[AGE]
The lacquer painting has a history of 80 years old	[Non-sensitive]
I've suffered depression and other mental probs since my teens	[MENTAL HEALTH]
Mental illness can also be an invisible disability	[NON-SENSITIVE]

TABLE 3. Size of the dataset used for experiments.

	SPEDAC 1	SPEDAC 2	SPEDAC 3
Training set	7611	3695	3893
Validation set	846	411	556
Test set	2218	1027	1112

TABLE 4. Krippendorff's (α) between gold and single annotation.

	Ann. 1	Ann. 2	Ann. 3	Ann. 4
SPEDAC 1	0.84	0.82	0.68	0.68
SPEDAC 2	0.82	0.84	0.92	/
SPEDAC 3	0.94	0.90	0.84	0.88

3) SPEDAC 3

Identification of fine-grained PDC within sensitive sentences. As we just said, SPEDAC 2 represents the specific PDCs considered in a balanced way. These sentences, along with other examples, are also labeled as specific categories (5,562 in total) and constitute the fine-grained SPEDAC dataset. The specific density of each fine-grained sensitive category is presented in Table 5.

4) INTER-ANNOTATOR AGREEMENT

To measure the goodness of our annotations, we asked a group of linguists to annotate a sample from each corpus. The basis given to them for annotation was the taxonomy of DPV-PD.

1) SPEDAC 1: we asked four annotators to binary classify 100 sentences as sensitive or non-sensitive. Giving the taxonomy as a reference, it was specified not to mark as sensitive only the sentences containing PII but to follow a more extensive definition of personal information that takes into consideration all the PDCs listed in the provided taxonomy;
2) SPEDAC 2: we asked three annotators to classify 150 sentences over the 5 macro-categories of the PDCs. In addition to the taxonomy, a detailed definition of the 5 macro-categories was provided, with examples of PDCs included in each group;
3) SPEDAC 3: because the specific PDCs are numerous, we have limited the task to the validation of our first labeling on 50 sentences. We received a contribution from four annotators. They were asked to compare the specific PDCs with which they found the sentences labeled with the definition given in the DPV-PD.

Sentences were randomly selected, balancing the number of different labels on SPEDAC 1 and SPEDAC 2. We measured the score agreement by aggregating the original annotation with the others using Krippendorff's alpha (α) coefficient [56]. Krippendorff's α expresses the score in terms of disagreement and is recommended if there are three or more annotators, attenuating the statistical effects of sample low-size datasets and ignoring missing data that may be present in collaborative work. Values range from 0 to 1, where 0 indicates perfect disagreement and 1 indicates perfect agreement. (α) $\geq$ .800 is usually considered a high agreement, while an acceptable agreement is considered in Krippendorff [57] as .667 $\geq$ (α) $\geq$ .800, even if the various proposals of the scholars highlight the arbitrary character of the reference thresholds [58].

The Krippendorff's (α) is 0.73 for SPEDAC 1, 0.82 for SPEDAC 2, and 0.87 for SPEDAC 3 respectively. The score obtained by comparing the gold annotation with the annotation or validation of each annotator was also measured. The results are summarized in Table 4. It might be unusual to see a higher percentage of agreement in the second task, where there are more labels. In SPEDAC 1 the sentences that reported a high rate of disagreement are mostly (i) ambiguous sentences, in which potentially sensitive personal data is expressed, as well as the relationship with a subject, but appear within a non-sensitive context (e.g., a fictitious example to explain a concept); (ii) sentences in which potentially sensitive personal data appears but the subject is not uniquely identifiable (often an unspecified group of people); (iii) sensitive sentences presenting specific personal data not identified by annotators, e.g., *House Owned*. On the other hand, despite obtaining an 1almost perfect' agreement score, SPEDAC 2 and SPEDAC 3 can sometimes present sentences that are potentially multi-labeled.

5) DATASET SPLIT

Each dataset was randomly divided into three parts for the experimental process: 70% training set, 10% development set, and 20% test set (see Table 3).

The distribution of labels in the training, validation and test sets of SPEDAC 1 and SPEDAC 2 can be observed in Table 6 and Table 7.

C. MODELS

We dedicate this paragraph to the description of the computational models used to conduct classification experiments on the different tasks offered by SPEDAC.

1) BASELINE

The baseline was calculated using the zero rate (ZeroR) classifier. This method draws the most-frequent baseline by roughly classifying all instances as corresponding to the most frequent class. It was included as a pure indicator of minimal rough classification results.

TABLE 5. Listed labels in SPEDAC 3 and PDCs included.

Label	PDCs	% dataset
[Age]	AGE, AGE EXACT, AGE RANGE, BIRTH DATE, BIRTH PLACE	1.87%
[Apartment Owned]	APARTMENT OWNED	1.74%
[CarOwned]	CAR OWNED	1.51%
[Country]	COUNTRY	1.71%
[Credit]	CREDIT	1.73%
[Criminal]	CRIMINAL,CRIMINAL CHARGE, CRIMINAL CONVICTION, CRIMINAL PARDON, CRIMINAL OFFENSE	0.25%
[Dialect]	DIALECT	2.03%
[Disability]	DISABILITY	1.80%
[Divorce]	DIVORCE	2.62%
[Drug Test Result]	DRUG TEST RESULT	2.45%
[Employment History]	EMPLOYMENT HISTORY	2.82%
[Ethnicity and Ethnic Origin]	ETHNICITY, NATIONALITY	1.85%
[Family and Family Structure]	FAMILY, FAMILY STRUCTURE	3.74%
[Family Health History]	FAMILY HEALTH HISTORY	2.16%
[Favorite]	FAVORITE	2.41%
[Favorite Color]	FAVORITE COLOR	1.28%
[Favorite Food]	FAVORITE FOOD	1.76%
[Favorite Music]	FAVORITE MUSIC	1.56%
[Fetish]	FETISH	1.19%
[Gender]	GENDER	1.76%
[Hair Color]	HAIR COLOR	1.55%
[Health]	HEALTH, HEALTH RECORD, MEDICAL HEALTH	1.94%
[Health History]	HEALTH HISTORY, INDIVIDUAL HEALTH HISTORY	1.78%
[Height]	HEIGHT	1.83%
[House Owned]	HOUSE OWNED	1.47%
[Income Bracket]	INCOME BRACKET	0.74%
[Job]	JOB	1.31%
[Language]	LANGUAGE	1.83%
[Location]	LOCATION, GEOGRAPHIC, DEMOGRAPHIC	0.86%
[Marital Status]	MARITAL STATUS	1.60%
[Marriage]	MARRIAGE	1.98%
[Mental Health]	MENTAL HEALTH	1.67%
[Name]	NAME	1.46%
[Offspring]	OFFSPRING	2.01%
[Ownership]	OWNERSHIP, PERSONAL POSSESSION, PERSONAL DOCUMENTS	1.83%
[Parent]	PARENT	1.26%
[Physical Characteristic and Trait]	PHYSICAL CHARACTERISTIC, PHYSICAL TRAIT	1.51%
[Physical Health]	PHYSICAL HEALTH	1.55%
[Piercing]	PIERCING	1.33%
[Political Affiliation]	POLITICAL AFFILIATION, POLITICAL OPINION	1.35%
[Prescription]	PRESCRIPTION	1.49%
[Privacy Preference]	PRIVACY PREFERENCE	1.01%
[Proclivitie]	PROCLIVITIE	1.51%
[Professional]	PROFESSIONAL,CURRENT EMP., PAST EMP., WORK ENVIRONMENT	1.31%
[Professional Certification]	PROFESSIONAL CERTIFICATION	1.62%
[Professional Evaluation]	PROFESSIONAL EVALUATION, PERFORMANCE AT WORK, DISCIPLINARY ACTION	0.16%
[ProfessionalInterview]	PROFESSIONAL INTERVIEW	1.11%
[Race]	RACE	1.55%
[Reference]	REFERENCE	1.67%
[Relationship]	RELATIONSHIP	1.69%
[Religion]	RELIGION	1.82%
[Salary]	SALARY	0.49%
[School]	SCHOOL, EDUCATION, EDUCATION EXPERIENCE, EDUCATION QUALIFICATION	1.74%
[Sexual]	SEXUAL	1.98%
[Sexual History]	SEXUAL HISTORY	1.82%
[Sexual Preference]	SEXUAL PREFERENCE	1.82%
[Sibling]	SIBLING	1.96%
[Skin Tone]	SKIN TONE	1.01%
[Tattoo]	TATTOO	1.60%
[Weight]	WEIGHT	1.82%
[Work History]	WORK HISTORY	1.73%

TABLE 6. Label distribution in SPEDAC 1.

	Train	% Train	Val	% Val	Test	% Test
Non-sens	3790	49.80%	405	47.87%	1086	48.96%
Sens	3821	50.20%	441	52.13%	1132	51.04%

TABLE 7. Label distribution in SPEDAC 2.

	Train	% Train	Val	% Val	Test	% Test
Special Data	979	26.49%	103	25.06%	274	26.68%
Financial and Tracking	468	12.67%	59	14.36%	122	11.88%
Social	1100	29.77%	137	33.33%	328	31.94%
Internal	321	8.69%	30	7.30%	83	8.08%
External	827	22.38%	82	19.95%	220	21.42%

2) k-NEAREST NEIGHBORS (*k*-NN)

k-NN is an algorithm used both for classification and regression, which is based on the similar characteristics of neighboring features [59]. The *k*-NN classifier uses instance-based learning, it does not build a general internal model, but stores instances of the training data. An instance is classified based on the plurality vote of its closest neighbors. The data class that has the greatest number of representatives within the closest neighbors to the instance is the predicted class. The number of neighbors to be considered is a parameter of the model to be established (k). In particular, for binary and multiclass classification, the number of neighbors should be odd. We trained the model implemented in sklearn [60]: KNeighborsClassifier,[1] where the optimal choice of the value k is highly data-dependent (generally, a larger k can reduce the noise, but makes the classification boundaries less distinct). The time complexity of the model is defined - following the Big O notation [61] - by the product of k=number of neighbors; d=number of data points; and n=number of neurons/data dimensionality. The time complexity of the models used is summarized in Table 8.

3) SUPPORT VECTOR MACHINES

Support vector machines (SVMs) are another classic algorithm [62] capable of building both binary and multiclass classifiers. SVMs use tagged data to define a hyperplane in which they map training examples, in an attempt to maximize the gap between categories. New examples are classified based on where they are mapped. For multiclass classification, the same principle is used after breaking down the classification problem into smaller subproblems, all of which are binary classification problems. A LIBSVM linear model was used from sklearn [60] for our experiments.[2] The model is recommended for sets that are not too wide: the fit time scales at least quadratically with the number of samples and therefore excludes its use on large datasets. The time complexity of SVMs is calculated with an exponent of 3 [63].

4) LOGISTIC REGRESSION

Logistic regression (LR) is a regression model implemented for binary and multiclass classification problems [64]. The model establishes the probability of identifying the value of the dependent variable by analyzing the attributes of the input and processing a weight distribution. The probability of belonging to a sample was calculated for each class. We used the LR model implemented in sklearn [60] for the experiments.[3] This implementation can fit the binary, One-vs-Rest, or multinomial logistic regression with optional penalty terms ($l1$,$l2$), or Elastic-Net regularization (by default). The time complexity is a product of data dimensionality and the number of data inputs.

5) TRANSFORMER-BASED LANGUAGE MODELS: RoBERTa AND DeBERTa

We adopted the RoBERTa (Robustly Optimized BERT Pre-training Approach) and DeBERTa (Decoding-enhanced BERT with disentangled attention) transformer architecture [30], [65]. RoBERTa has been proven to perform well in different NLP tasks, including classification [66], [67], [68]. DeBERTa models seem to perform consistently better on a wide range of NLP tasks even if trained on half of the training data [65]. RoBERTa and DeBERTa are both an extension of the Bidirectional Encoder Representations from Transformers (BERT) [69]. BERT uses two bidirectional training strategies: the masked language model (MLM), which deals with the relationship between words, and the next predictive sentence (NPS) to predict the relationship between sentences. BERT's architecture is composed of a tokenizer (WordPiece) and a large stack of transformers, which is provided with the input for training. The BERT-Base model consists of a 12-layer transformer, whereas the BERT-Large of a 24-layer. RoBERTa has almost the same architecture as BERT model, but uses a byte version of byte-pair encoding (BPE) as a tokenizer and is pretrained with the MLM task (without the NPS task). It optimizes some hyperparameters for BERT, e.g., longer training time, larger training data, larger batch size, larger vocabulary size, and dynamic masking. DeBERTa improves the BERT and RoBERTa models by adding two novel techniques. First, a disentangled attention mechanism uses two vectors to encode and separate the content and position of a word. Second, an enhanced mask decoder can predict both the relative and absolute position of words, while the previous models took into account only one of them.

We used the RoBERTa-base and DeBERTa-base models with pre-trained weights [70], [71], and 768 hidden dimensions. Time complexity is a product between n with an exponent of 2 and d considered per layer [28]. The additional computational complexity of DeBERTa is $O(k * n * d)$ due to the calculation of the additional position-to-content and

[1]The model implemented can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (last access December 05, 2022)

[2]The model implemented can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (last access December 05, 2022)

[3]The model implemented can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (last access December 05, 2022)

TABLE 8. Time complexity of the models following Big O notation [61].

Model	Time complexity	Layers and Parameters
k-NN	$O(k * n * d)$	/
SVMs	$O(n^3)$	/
LR	$O(n * d)$	/
LaBSE	$O(n^2 * d)$ per layer	12-layers, 471M
RoBERTa	$O(n^2 * d)$ per layer	12-layers, 125M
DeBERTa	$O(n^2 * d)$ per layer	12-layers, 98M

content-to-position attention scores. This increases the computational cost of RoBERTa by 30% [65].

6) SENTENCE-TRANSFORMERS LANGUAGE MODELS: LaBSE

The architecture of a sentence-transformer model is made of two main layers. The first layer is a transformer model with a fixed length of 768 dimensions that outputs contextualized word embeddings for all input tokens. The second layer is a pooling layer that averages the embeddings generated by the model giving a fixed length vector [72]. LaBSE (Language-agnostic BERT Sentence Embedding) [73] is a multilingual sentence embedding model for more than 109 languages, originally trained and optimized to produce similar representations exclusively for bilingual sentence pairs that are translations of each other. Thanks to its specific training, the model achieves state-of-the-art performance on bilingual retrieval/mining tasks. Multilingual sentence embedding models produce representations that are suitable to be compared with simple cosine similarity also on the same language [73], [74]. The study aims to investigate the use of a LaBSE model on the classification task.

The used encoder architecture follows the BERT-Base model, with 12 hidden layers and 768 per-position hidden units. Sentence embeddings are extracted from the last transformer block [75].

IV. EXPERIMENTAL SETUP

The datasets created and described in Section III-B were used first for the experiments conducted with the transformer models and then, for comparative purposes, with the other models.

A. EXPERIMENT 1

Dataset. Identification of sensitive sentences and exclusion of non-sensitive sentences. In particular, as described above, we built an adversarial dataset of sentences with non-sensitive content that is particularly competitive with the sensitive content dataset. The sentences in both datasets contain the same linguistic patterns; what differentiates a sensitive sentence from a non-sensitive one is the context in which it occurs. The same datasets were used to perform all the experiments. The subdivision, as described in Tables 3, 6, occurred randomly only once and the derived datasets were used to train and test all the models.

Preprocessing, features and parameters. First, the data were preprocessed and cleaned. The preprocessing process includes the tokenization of sentences, lemmatization, conversion of each token into a lowercase, removal of spaces, stop words, and punctuation. Feature extraction on the training set was performed using the scikit learn feature extraction from the text modules. The features are not domain-dependent but English language-dependent, and are the following:

- whether a token starts and ends a sentence;
- the length of the sentences in tokens;
- bag-of-words (BOW) vectors (ngram range=1,1) using the SPaCy CountVectorizer function [60].

Preprocessing and feature extraction using SpaCY are common for all three classic machine learning models implemented (kNN, SVM, and LR).

The model parameters were set up and tuned on the SPEDAC validation set as follows:

- for the k-NN model, we considered the 3 closest neighbors (k=3);
- for the SVM model, we used default parameters to set up a linear kernel;
- for the LR model, default parameters have been used;
- for the transformer models, we set a stack with a dropout level of 0.3, and a randomly initialized linear transformation level above the model. The maximum sequence length was set to 256, and the training lot size was set to 8. For the model optimization, we used the AdamW optimizer [76] with a learning rate of $1e$-5. The performance was evaluated based on the loss of the binary cross-entropy. After 3 epochs, the model reports a training accuracy epoch beyond 0.90 on the validation set.

B. EXPERIMENT 2

Dataset. Identification of which type of sensitive data the sentence presents, related to its macro-category. Once the sensitive sentences have been identified (layer 1, Fig. 2), they are analyzed by the multiclass model, which labels them according to the 5 macro-categories on which they are trained (layer 2, Fig. 2).

Even in this case, the same datasets were used to run all the experiments and their subdivisions are described in Tables 3, 7.

Preprocessing, features and parameters. The preprocessing process and feature extraction are the same as in the first experiment.

The parameters of the models, set up and tuned on the SPEDAC validation set, are as follows:

- for the k-NN model, we considered the 3 closest neighbors (k=3);
- for the SVM model, the multiclass classification strategy used follows the One-vs-One (OvO) scheme, which involves breaking down the multiclass classification into a binary classification problem for each pair of classes;
- for the LR model, this case, for the multiclass classification we used the One-vs-Rest (OvR) scheme, which

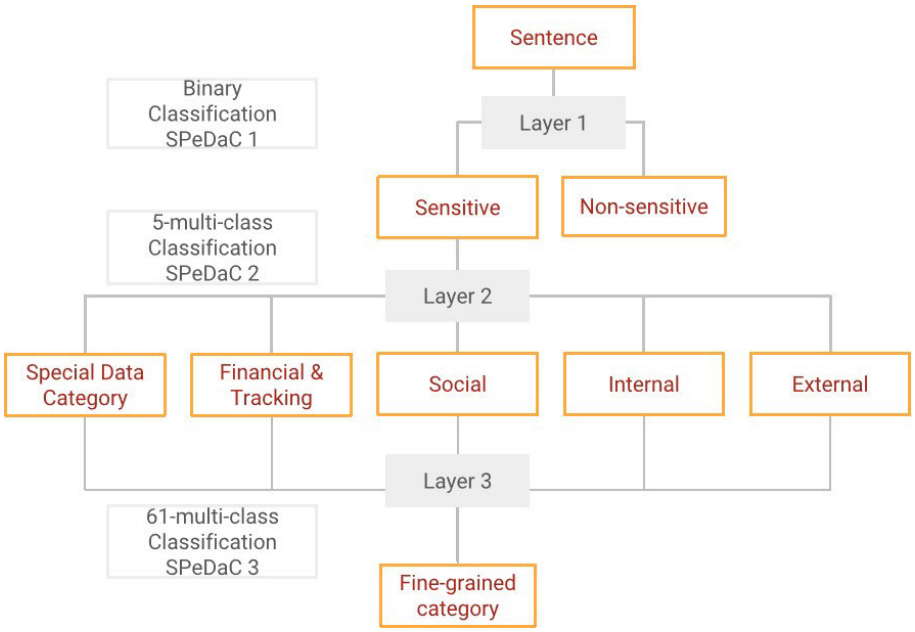


FIGURE 2. Flow of sensitive detection model.

divides multiclass classification into a binary classification problem by class;
- for the transformer models, the setting is the same as for SPEDAC 1 and likewise reports a training accuracy epoch beyond 0.90 on the validation set.

C. EXPERIMENT 3

Dataset. Identification of the type of fine-grained PDC in a sentence. This involves a multiclass classification task with 61 labels and a small amount of training data for each PDCs.

Preprocessing, features and parameters. The models used were the same as those in the second experiment with the following differences:
- for the baseline of SPEDAC 3, the 61 labels were traced to the macro-category and the most-frequent baseline was calculated by tracing all the test sentences to the most frequent macro-category;
- for the k-NN model, 5 closest neighbors have been used (k=5);
- to improve the LR results, a liblinear solver with penalty *l1* was applied;
- to improve the results of the transformer models, a category regularization with a label smoothing technique was introduced [77] and the number of epochs in training was increased to 15.

V. RESULTS

The model predictions were evaluated in terms of accuracy.

EXPERIMENT 1

The results of the first and second experiments on SPEDAC 1 are listed in Table 9. As can be seen, RoBERTa reports the best results compared with the other models for the binary classification task for sensitive and non-sensitive sentence identification even if DeBERTa and LaBSE both report very high results as well. SPEDAC 1, as described in Section III-B, is composed of sensitive and non-sensitive sentences that have the same linguistic patterns, which acquire sensitivity or not depending on the context. If the discriminant of sensitive and non-sensitive sentences in the dataset often consists of contextual elements, given the occurrence of the same linguistic patterns, the transformer context-aware models turn out to be the most suitable for the task.

TABLE 9. Accuracy results on SPEDAC 1 and SPEDAC 2.

	SPEDAC 1	SPEDAC 2
Baseline	51.04%	31.93%
k-NN	68.62%	63.78%
SVM	93.15%	92.30%
LR	92.60%	92.50%
LABSE	98.15%	94.84%
RoBERTa	**98.20%**	94.94%
DeBERTa	98.11%	**95.81%**

EXPERIMENT 2

The results of the first and second experiments on SPEDAC 2 are listed in Table 9. In the multi-class classification of SPEDAC 2, where the problem of ambiguity is less evident, the results obtained with the other models are more promising. The DeBERTa model outperforms the other models in all cases, and the RoBERTa model surpasses the LR performance by 2.44%. It is interesting to observe how LaBSE, a very promising model for multilingual sentence similarity, does not achieve the best results for the classification task when compared to the other transformer models. Probably

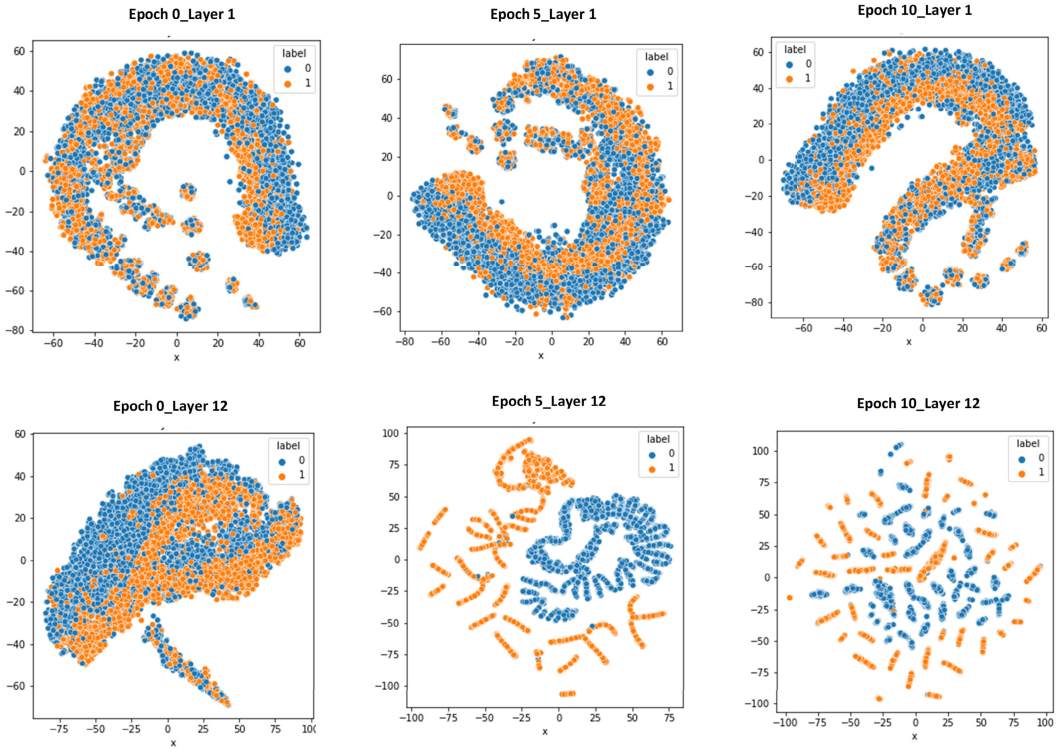


FIGURE 3. RoBERTa embeddings t-SNE visualization during 2-class fine-tuning (perplexity=30).

TABLE 10. Accuracy results on SPEDAC 3: a new benchmark.

	SPEDAC 3
Baseline	32.25%
k-NN	35.30%
SVM	57.59%
LR	75.74%
LABSE	77.09%
RoBERTa	77.18%
DeBERTa	**77.63%**

because it was trained to detect similar sentences in different languages.

EXPERIMENT 3

The performances of the models of the third experiment on SPEDAC 3 are presented in Table 10. The results, which differ significantly between the models in terms of percentage accuracy offer valid results for a benchmark on the SPEDAC 3.

A. RESULTS ANALYSIS

Fig. 3 and Fig. 4 show a t-SNE visualization [78] of the RoBERTa embeddings during the fine-tuning of training data. The first and last hidden layers of the transformer network are reported. During the validation stage, the weights of the model are not updated. From visualizations, it can be seen that for both tasks, already after epoch 5, the embeddings are distinctly clustered.

To better understand the behavior of the model on SPEDAC 1 and SPEDAC 2, we report the results in terms of

TABLE 11. Confusion matrix sensitiveness classification on SPEDAC 1.

Actual Class	RoBERTa		DeBERTA		SVM	
	Non-sens	**Sens**	**Non-sens**	**Sens**	**Non-sens**	**Sens**
Non-sens	**1102**	30	**1104**	28	**1036**	96
Sens	10	**1076**	14	**1072**	56	**1030**

the accuracy for each classification category taking RoBERTa and DeBERTa as the transformer models that obtain the best results and SVMs for the ML comparison methods (see Table 11,12). By analyzing the errors through confusion matrices, we see how the RoBERTa and DeBERTa models obtain the best performance for each category without significant differences; there are no particularly critical categories to classify.

However, it should be noted that, in terms of time complexity (Table 8), the ML models report significantly lower values than the transformer ones.

Concerning the specific experiments we can make the following considerations:

EXPERIMENT 1

The models mostly failed to identify non-sensitive sentences, although RoBERTa and DeBERTa are considerably more accurate. By analyzing errors, many sentences are misidentified as sensitive presumably because of the high rate of ambiguity they present. Errors are caused by the presence of expressions and keywords related to health or profession and

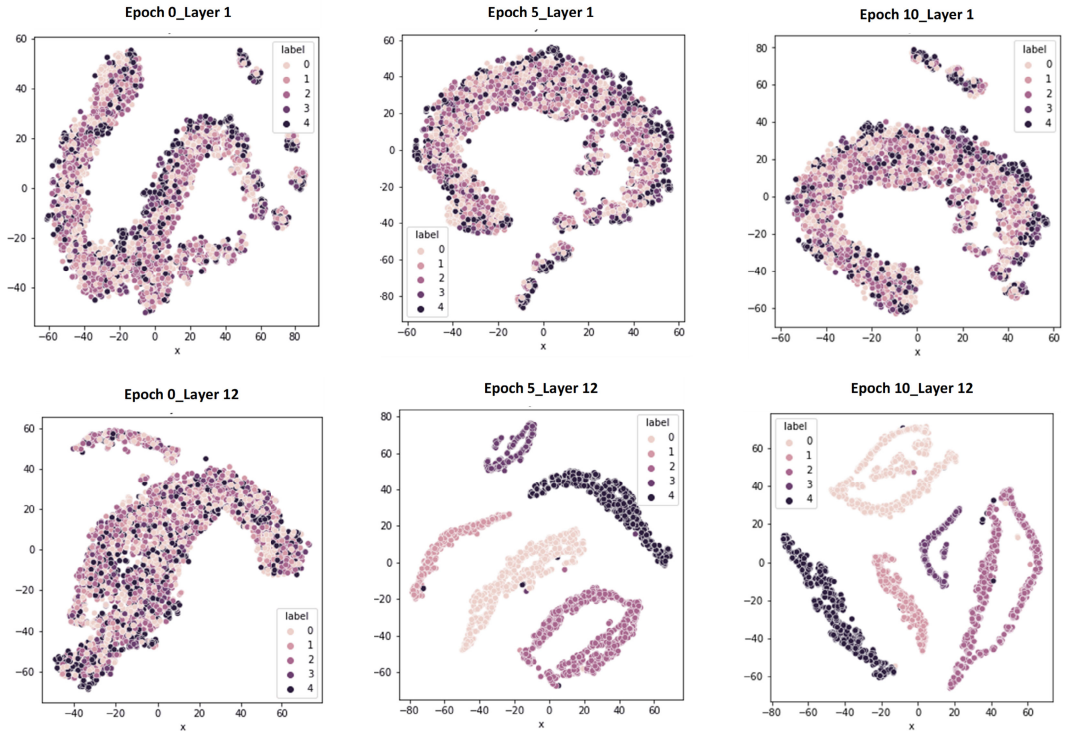


FIGURE 4. RoBERTa embeddings t-SNE visualization during 5-class fine-tuning (perplexity=30).

TABLE 12. Confusion matrix macro-categories classification on SPEDAC 2.

Actual Class	RoBERTa					DeBERTa					SVM				
	Spec	**Fin**	**Soc**	**Int**	**Ext**	**Spec**	**Fin**	**Soc**	**Int**	**Ext**	**Spec**	**Fin**	**Soc**	**Int**	**Ext**
Spec	**258**	0	10	0	6	**259**	0	4	1	10	**248**	1	14	0	11
Fin	1	**115**	6	0	0	0	**120**	1	0	1	1	**114**	6	0	1
Soc	4	2	**319**	2	1	8	3	**312**	3	2	7	7	**305**	1	8
Int	1	0	1	**81**	0	1	0	0	**82**	0	0	0	0	**81**	2
Ext	7	1	10	0	**202**	6	1	2	0	**212**	12	2	6	0	**200**

the model in these cases is unable to discriminate assumptions or hopes useful to exclude the sensitivity of the sentence e.g., '*I had great hopes of being an air hostess so that i could travel to so many places than I heard about a plane crushed and that kind of threw me off the idea*'. However, it is important to note that this is not a systematic error: RoBERTa at the same time classifies as non-sensitive sentences where the profession of the subject is only a guess e.g., '*I 'm supposed to be a movie critic, and yet I keep hearing about these great new movies I've never seen.*'. This leads us to consider that cases of ambiguity can be addressed by adding training sentences to represent them.

EXPERIMENT 2

Contrary to what one might assume, the category with fewer training examples (*Internal*) achieved a high accuracy score. Indeed, the macro category has fewer examples, but at the same time has fewer specific categories of personal data that represent it. All the specific PDCs belonging to the macro PDC *Internal* refer to personal preferences (*Preference*, *Favorite Color*, *Favorite Music*) and are therefore well identified by the model. RoBERTa mainly mistakes the macro PDC *External* for the *Social* and the category is generally confused with the *Special Data* category. Furthermore, it can be seen that the models confuse some sentences classified as *Special Data* or *External* with the *Social* category. This can be explained by the fact that some sentences contained more than one PDC. Therefore, they would need double-label and classification.

EXPERIMENT 3

As for the SPEDAC 2 experiment, the models that achieved the highest performance were the RoBERTa and DeBERTa transformer models and the LR-based models.

By conducting an error analysis on the predictions of the models, we identified systematic confusion between targets and predictions, highlighting the errors that exceeded 20% (see Table 13). The confusing labels often belong to the same

TABLE 13. % Errors ≥ 20% in SPEDAC 3 (RoBERTa, DeBERTa and LR models). When '/' appears, it means a % of error < 20.

Target	Pred	% Error RoB	% Error DeB	% Error LR
Ethnicity and Ethnic Origin	Skin Color	21.40%	21.40%	21.40%
Family Health History	Drug Test Result	20%	20%	20%
Favorite Food	Favorite	/	/	28.50%
Location	Country	33.30%	20%	40%
Health History	Health	24%	/	/
Mental Health	Health	20%	26.70%	20%
Physical Characteristic and Trait	Hair Color	36.80%	26%	31.50%
Professional Evaluation	Reference	25%	/	50%
Reference	Employment History	20%	/	/
Reference	Professional Interview	/	20%	20%
Salary	Parent	25%	25%	25%
Salary	Credit	25%	/	/
Salary	Family and Family Structure	/	/	25%
School	Professional Certification	31.20%	/	/
Sexual	Proclivitie	31.80%	31.80%	/
Sexual History	Sexual	/	/	28%
Work History	Employment History	47.60%	38%	61.90%

macro-category and present similarities in terms of keywords and linguistic patterns.

Another significant problem that emerges from the error analysis concerns sentences that contain more than one sensitive data item which would require multi-category labeling. '*Nancy and I were married in 1977 and we lived for nearly 30 years in the Duveneck school area*' is a sentence that reveals sensitive information that can be traced back to two categories: *Marital Status* and *Location*. In future works, it is expected that this problem will be solved by the span-based labeling of SPEDAC.

VI. CONCLUSION AND FUTURE WORK

In this study, we investigated the task of automatic sensitive data identification and classification, based on our work on personal data categories, which has not been explored in the literature. To do this, we created labeled datasets. The SPEDAC corpora were evaluated by comparing machine learning algorithms, including the transformer models, with which we achieved the best results. An accuracy of over 90% was achieved in the classification of sensitive and non-sensitive sentences (SPEDAC 1) and the discrimination of the 5 macro-categories of personal data. Lower results (<80% acc.) are achieved in the 61-class classification of SPEDAC 3. This dataset can be used as a valid benchmark for future studies.

First, the most important goal achieved in this work concerns the creation of the SPEDAC labeled datasets for the task of automatic identification of personal data, based on the taxonomy of the DPV. The datasets constitute an available resource and a benchmark for the task, which is currently not present in the literature. Future work foresees the expansion of the SPEDAC corpora both quantitatively and multilingually. In particular, we would like to consider the Italian language. Therefore, as anticipated and based on the error analysis conducted, it would be very useful to label SPEDAC at a finer level than the sentence-based one, labeling multiple PDCs on every single sentence. We assume token-level labeling following the BIO encoding format.

Second, to evaluate SPEDAC, we explored a model based on deep learning for the identification of sentences with sensitive content and the classification of the personal data macro-categories present in them. The hypothesis that pre-trained transformer networks based on multi-head attention modules can perform classification tasks whose labels are highly context-dependent has been confirmed by the results. Indeed, binary and 5-label classification tasks conducted on the BERT extension, DeBERTa, report extremely high accuracy results and appear to be the best especially when compared to different automatic learning models (k-NN, SVM, and LR).

However, the deep learning approach does not seem to achieve excellent results when there are few training data and many classification labels, as in the case of SPEDAC 3, although model adaptation techniques (e.g., label smoothing) can improve them. Combining a logical-symbolic approach that requires little or no training data could be an interesting solution to explore [79].

In any case, the comparison with the state-of-the-art when implementing different identification techniques is always very difficult, because of the lack of shared resources and benchmarks. The SPEDAC resource contributes in this sense. The datasets can be shared under an ethical disclosure agreement and used to evaluate other identification and classification models for PDCs.

To conclude, the SID task we have addressed, which - as aforementioned - is a subpart of the DLD, helps to improve the DLP systems. The resource and the results intercept an industrial interest. Future works could also explore and test the model to search for and identify sensitive information in structured data. Finally, SPEDAC could be extended to identify other sensitive data categories at high risk of DLD e.g., passwords left in scripts and software codes.

VII. ETHICAL DISCLOSURE

The automatic processing of sensitive data implies a necessary reflection on the ethical aspects and improper uses derived from this type of research [80], [81]. The created dataset presents publicly available texts, labeled by categories of sensitive data but in no way attributable to identifiable subjects. This dataset simulates the contexts of sensitivity but is not sensitive. Nevertheless, the trained model can certainly be used for malicious purposes, in contrast to what we pursue. To avoid this possibility, we have bound the download of SPEDAC to the prior signing of an agreement by the user that establishes ethical research purposes.

REFERENCES

[1] M. Larson, N. Oostdijk, and F. Zuiderveen Borgesius, *Not Directly Stated, Not Explicitly Stored: Conversational Agents and the Privacy Threat of Implicit Information*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 388–391, doi: 10.1145/3450614.3463601.

[2] K. Adhikari and R. Panda, "Users information privacy concerns and privacy protection behaviors in social networks," *J. Global Marketing*, vol. 31, no. 2, pp. 96–110, Jan. 2018, doi: 10.1080/08911762.2017.1412552.

[3] I. Hendrickx, J. Van Waterschoot, A. Khan, L. T. Bosch, C. Cucchiarini, and H. Strik, "Take back control: User privacy and transparency concerns in personalized conversational agents," in *Proc. IUI Workshops*, 2021, pp. 1–7.

[4] *Cost of a Data Breach Report*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.ibm.com/downloads/cas/RDEQK07R

[5] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," 2017, *arXiv:1707.02919*.

[6] *Eu General Data Protection Regulation (EU-GDPR)*. Accessed: Jan. 23, 2023. [Online]. Available: https://gdpr.eu/

[7] *Eu General Data Protection Regulation (EU-GDPR)*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.privacy-regulation.eu/en/r6.htm

[8] *Eu General Data Protection Regulation (EU-GDPR)*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.privacy-regulation.eu/en/4.htm

[9] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*. New York, NY, USA: Association for Computing Machinery, 2000, pp. 439–450, doi: 10.1145/342009.335438.

[10] E. E. Özkoç, "Privacy preserving data mining," in *Data Mining*, C. Thomas, Ed. Rijeka, Croatia: IntechOpen, 2021, ch. 3, doi: 10.5772/intechopen.99224.

[11] P. Cheng, J. F. Roddick, S.-C. Chu, and C.-W. Lin, "Privacy preservation through a greedy, distortion-based rule-hiding method," *Int. J. Speech Technol.*, vol. 44, no. 2, pp. 295–306, May 2015, doi: 10.1007/s10489-015-0671-0.

[12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 99–106, doi: 10.1109/ICDM.2003.1250908.

[13] Y. Xiao, L. Xiao, X. Lu, H. Zhang, S. Yu, and H. V. Poor, "Deep-reinforcement-learning-based user profile perturbation for privacy-aware recommendation," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4560–4568, Mar. 2021, doi: 10.1109/JIOT.2020.3027586.

[14] R. X. Lu, H. Zhu, J. K. Liu, J. Shao, and X. Liu, "Toward efficient and privacy-preserving computing in big data era," *IEEE Netw.*, vol. 28, no. 4, pp. 46–50, Jul./Aug. 2014, doi: 10.1109/MNET.2014.6863131.

[15] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system," *Inf. Sci.*, vol. 479, pp. 567–592, Apr. 2019, doi: 10.1016/j.ins.2018.02.005.

[16] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002, doi: 10.1142/S0218488502001648.

[17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond K-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, 2006, p. 24, doi: 10.1109/ICDE.2006.1.

[18] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond K-anonymity and L-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[19] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[20] J. C.-W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, and M. Aloqaily, "Privacy-preserving multiobjective sanitization model in 6G IoT environments," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5340–5349, Apr. 2021, doi: 10.1109/JIOT.2020.3032896.

[21] C.-W. Lin, T.-P. Hong, and H.-C. Hsu, "Reducing side effects of hiding sensitive itemsets in privacy preserving data mining," *Sci. World J.*, vol. 2014, pp. 1–12, Jan. 2014, doi: 10.1155/2014/235837.

[22] J. C.-W. Lin, T.-Y. Wu, P. Fournier-Viger, G. Lin, J. Zhan, and M. Voznak, "Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining," *Eng. Appl. Artif. Intell.*, vol. 55, pp. 269–284, Oct. 2016, doi: 10.1016/j.engappai.2016.07.003.

[23] J. C.-W. Lin, P. Fournier-Viger, L. Wu, W. Gan, Y. Djenouri, and J. Zhang, "PPSF: An open-source privacy-preserving and security mining framework," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 1459–1463, doi: 10.1109/ICDMW.2018.00208.

[24] M. Hart, P. Manadhata, and R. Johnson, "Text classification for data loss prevention," in *Privacy Enhancing Technologies* (Lecture Notes in Computer Science), vol. 6794, S. Fischer-Hübner and N. Hopper, Eds. Berlin, Germany: Springer, 2011, pp. 18–37, doi: 10.1007/978-3-642-22263-4_2.

[25] *Google De-Identify Sensitive Data Tool*. Accessed: Jan. 23, 2023. [Online]. Available: https://cloud.google.com/dlp/docs/ deidentify-sensitive-data#api_overview

[26] *IBM Discover Sensitive Data Tool*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.ibm.com/docs/en/guardium/10.6?topic=discover-sensitive-data

[27] *Microsoft PII Detection Tool, Azure Cognitive Service*. Accessed: Jan. 23, 2023. [Online]. Available: https://docs.microsoft.com/ en-us/azure/cognitive-services/language-service/personally-identifiable-information/

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Proc. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, pp. 1–11.

[29] J. Neerbek, M. Eskildsen, P. Dolog, and I. Assent, "A real-world data resource of complex sensitive sentences based on documents from the monsanto trial," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 1258–1267.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[31] M. Dias, J. Boné, J. C. Ferreira, R. Ribeiro, and R. Maia, "Named entity recognition for sensitive data discovery in Portuguese," *Appl. Sci.*, vol. 10, no. 7, p. 2303, Mar. 2020, doi: 10.3390/app10072303.

[32] Y. Guo, J. Liu, W. Tang, and C. Huang, "Exsense: Extract sensitive information from unstructured data," *Comput. Secur.*, vol. 102, Mar. 2021, Art. no. 102156, doi: 10.1016/j.cose.2020.102156.

[33] A. G. Pablos, N. Perez, and M. Cuadros, "Sensitive data detection and classification in Spanish clinical text: Experiments with BERT," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 1–9.

[34] A. Genetu and T. Tegegne, "Designing sensitive personal information detection and classification model for amharic text," in *Proc. Int. Conf. Inf. Commun. Technol. Develop. Afr. (ICTDA)*, Nov. 2021, pp. 54–58, doi: 10.1109/ICT4DA53266.2021.9672227.

[35] J. Neerbek, "Sensitive information detection: Recursive neural networks for encoding context," Ph.D. dissertation, Dept. Comput. Sci., Aarhus Univ., Aarhus, Denmark, 2020.

[36] A. C. Islam, J. Walsh, and R. Greenstadt, "Privacy detective: Detecting private information and collective privacy behavior in a large social network," in *Proc. 13th Workshop Privacy Electron. Soc.*, Nov. 2014, pp. 35–46, doi: 10.1145/2665943.2665958.

[37] R. Chow, P. Golle, and J. Staddon, "Detecting privacy leaks using corpus-based association rules," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 893–901, doi: 10.1145/1401890.1401997.

[38] L. Geng, Y. You, Y. Wang, and H. Liu, "Privacy measures for free text documents: Bridging the gap between theory and practice," in *Trust, Privacy and Security in Digital Business* (Lecture Notes in Computer Science), vol. 6863, S. Furnell, C. Lambrinoudakis, and G. Pernul, Eds. Berlin, Germany: Springer, 2011, doi: 10.1007/978-3-642-22890-2_14.

[39] G. McDonald, C. Macdonald, and I. Ounis, "Enhancing sensitivity classification with semantic features using word embeddings," in *Advances in Information Retrieval*, J. M. Jose, C. Hauff, I. S. Altingovde, D. Song, D. Albakour, S. Watt, and J. Tait, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1145/3450614.3463601.

[40] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, D. Touretzky, Ed. 1989, pp. 1–9.

[41] G. Xu, C. Qi, H. Yu, S. Xu, C. Zhao, and J. Yuan, "Detecting sensitive information of unstructured text using convolutional neural network," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2019, pp. 474–479, doi: 10.1109/CyberC.2019.00087.

[42] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[43] Y. Lin, G. Xu, G. Xu, Y. Chen, and D. Sun, "Sensitive information detection based on convolution neural network and bi-directional LSTM," in *Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2020, pp. 1614–1621, doi: 10.1109/TrustCom50675.2020.00223.

[44] *Pastebin*. Accessed: Jan. 23, 2023. [Online]. Available: https://pastebin.com/

[45] W. Shi, R. Shea, S. Chen, C. Zhang, R. Jia, and Z. Yu, "Just fine-tune twice: Selective differential privacy for large language models," 2022, *arXiv:2204.07667*.

[46] *W3c*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.w3.org/

[47] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, E. Schlehahn, S. Steyskal, and R. Wenning, "Creating a vocabulary for data privacy," in *Proc. Move Meaningful Internet Syst., OTM Conf.*, H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. A. Ardagna, and R. Meersman, Eds. 2019, pp. 714–730, doi: 10.1007/978-3-030-33246-4_44.

[48] *Data Privacy Vocabulary (DPV)*. Accessed: Jan. 23, 2023. [Online]. Available: https://w3c.github.io/dpv/dpv/

[49] *DPV-Pd: Extended Personal Data Concepts for DPV*. Accessed: Jan. 23, 2023. [Online]. Available: https://w3c.github.io/dpv/dpv-pd/

[50] *Enron Email Dataset*. [Online]. Accessed: Jan. 23, 2023. Available: https://www.cs.cmu.edu/~enron/

[51] *Monsanto Papers*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.baumhedlundlaw.com/toxic-tort-law/monsanto-roundup-lawsuit/monsanto-papers/

[52] D. Sánchez and M. Batet, "C-sanitized: A privacy model for document redaction and sanitization," *J. Assoc. Inf. Sci. Technol.*, vol. 67, pp. 148–163, Jun. 2014, doi: 10.1002/asi.23363.

[53] B. Andreas, M. David, and W. Muiris, "Protecting personally identifiable information (PII) using tagging and persistence of PII," U.S. Patent 0 885 225, Jan. 5, 2021.

[54] M. Jakubícek, A. Kilgarriff, V. Kovár, P. Rychlý, and V. Suchomel, "The tenten corpus family," in *Proc. 7th Int. Corpus Linguistics Conf.*, 2013, pp. 125–127.

[55] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych, "The inception platform: Machine-assisted and knowledge-oriented interactive annotation," in *Proc. COLING*, 2018, pp. 1–5.

[56] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, no. 1, pp. 77–89, Apr. 2007, doi: 10.1080/19312450709336664.

[57] K. Krippendorff, "Reliability in content analysis: Some common misconceptions and recommendations," *Hum. Commun. Res.*, vol. 30, no. 3, pp. 411–433, Jul. 2004, doi: 10.1111/j.1468-2958.2004.tb00738.x.

[58] G. Gagliardi, "Inter-annotator agreement in linguistica: Una rassegna critica," in *Proc. Italian Conf. Comput. Linguistics*, Dec. 2018, pp. 206–212.

[59] L. Wang and X. Zhao, "Improved KNN classification algorithms research in text categorization," in *Proc. 2nd Int. Conf. Consum. Electron., Commun. Netw. (CECNet)*, Apr. 2012, pp. 1848–1852, doi: 10.1109/CECNet.2012.6201850.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "SciKit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[61] M. J. Kearns, "The computational complexity of machine learning," Ph.D. dissertation, Dept. Comput. Sci., Harvard Univ., Cambridge, MA, USA, 1989.

[62] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Oct. 1995, doi: 10.1007/BF00994018.

[63] A. Abdiansah and R. Wardoyo, "Time complexity analysis of support vector machines (SVM) in LibSVM," *Int. J. Comput. Appl.*, vol. 128, no. 3, pp. 28–34, Oct. 2015.

[64] E. Bisong, *Logistic Regression*. Berkeley, CA, USA: Apress, 2019, pp. 243–250, doi: 10.1007/978-1-4842-4470-8_20.

[65] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.

[66] J. Briskilal and C. N. Subalalitha, "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa," *Inf. Process. Manag.*, vol. 59, no. 1, Jan. 2022, Art. no. 102756, doi: 10.1016/j.ipm.2021.102756.

[67] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 1872–1897, 2020, doi: 10.1007/s11431-020-1647-3.

[68] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews," *Electron. Commerce Res.*, pp. 1–21, Apr. 2022, doi: 10.1007/s10660-022-09560-w.

[69] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chap. Assoc. Comput. Linguistics, Human Language*, Jun. 2019, pp. 1–16.

[70] *Roberta Base Model*. Accessed: Jan. 23, 2023. [Online]. Available: https://huggingface.co/roberta-base

[71] *Deberta Base Model*. Accessed: Jan. 23, 2023. [Online]. Available: https://huggingface.co/microsoft/deberta-base

[72] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.* Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

[73] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," 2020, *arXiv:2007.01852*.

[74] R. Tripodi, R. Blloshmi, and S. L. Sullam, "Evaluating multilingual sentence representation models in a real case scenario," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 2928–2939. [Online]. Available: https://aclanthology.org/2022.lrec-1.314

[75] *Labse Model*. Accessed: Jan. 23, 2023. [Online]. Available: https://huggingface.co/sentencetransformers/LaBSE

[76] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.

[77] R. Müller, S. Kornblith, and G. Hinton, *When Does Label Smoothing Help?*. Red Hook, NY, USA: Curran Associates, 2019.

[78] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

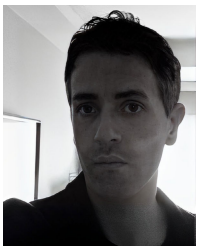
[79] G. Gambarelli and A. Gangemi, "PRIVAFRAME: A frame-based knowledge graph for sensitive personal data," *Big Data Cognit. Comput.*, vol. 6, no. 3, p. 90, 2022, doi: 10.3390/bdcc6030090.

[80] S. Suster, S. Tulkens, and W. Daelemans, "A short review of ethical challenges in clinical natural language processing," in *Proc. 1st ACL Workshop Ethics Natural Lang. Process.* Valencia, Spain: Association for Computational Linguistics, 2017, pp. 80–87, doi: 10.18653/v1/W17-1610.

[81] L. Weidinger et al., "Ethical and social risks of harm from language models," 2021, *arXiv:2112.04359*.

GAIA GAMBARELLI was born in Correggio, Reggio Emilia, Italy, in 1994. She received the M.S. degree in Italian studies, linguistics, and European literary cultures, with a thesis in computational linguistics, from the University of Bologna, in March 2019, where she is currently pursuing the Ph.D. degree in digital humanities with the Department of Classical Philology and Italian Studies (FICLIT). She has industrial work experience as an expert in conversational agents. Her Ph.D. project concerns the protection of sensitive data through textual automatic identification. Her main current research interests include natural language processing, semantics, and linguistics applied to privacy protection and conversational agents. Previously, she worked on automatic personality recognition, rhetoric, and the theory of argumentation.

ALDO GANGEMI is currently a Full Professor with the University of Bologna and the Director of the Institute for Cognitive Sciences and Technologies, Italian National Research Council, where he co-founded the Semantic Technology Laboratory (STLab), in 2008. He has published more than 250 papers in international peer-reviewed journals, conferences, and books (scholar H-index = 61) and sits as an EiC member or an EB member of international journals (semantic web, web semantics, and applied ontology), the Conference Chair (EKAW2008, WWW2015, and ESWC2018/9), has coordinated research teams in eight EU projects, and is the Scientific Coordinator of the H2020 SPICE Project. His research interests include semantic technologies as an integration of methods from knowledge engineering, semantic web, linked data, cognitive science, and natural language processing. His theoretical interests concentrate upon the representation and discovery of knowledge patterns across data, ontologies, natural language, and cognition, using hybrid symbolic/sub-symbolic methods, applications domains include cultural heritage, robotics, medicine, law, e-government, agriculture and fishery, and business. He is also a member of the Board of Directors at the IMT School for Advanced Studies Lucca.



ROCCO TRIPODI received the Ph.D. degree in computer science from the Ca' Foscari University of Venice, with a thesis titled "Evolutionary game theoretic models for natural language processing," in 2015. He was a Research Assistant and an Adjunct Professor at Ca' Foscari University, where he worked on lexical semantics and taught corpus linguistics, natural language processing, and digital text analysis. He worked as a Researcher in different laboratories, including Sapienza NLP, Sapienza University of Rome, and the European Centre for Living Technology (ECLT), Venice, and on different European projects, including ODYCCEUS, MOUSSE, and Polifonia. He is currently an Assistant Professor at the University of Bologna. His research interests include the areas of machine learning and natural language processing with a focus on lexical and sentence level semantics, learning models based on game theoretic principles, and the design, learning, and evolution of linguistic communication systems.

Open Access funding provided by 'Alma Mater Studiorum - Università di Bologna' within the CRUI CARE Agreement