Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

CULTURE LETTERARIE E FILOLOGICHE

Ciclo XXXV

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: L-LIN/01 - GLOTTOLOGIA E LINGUISTICA

DATA SENSITIVITY DETECTION IN CHAT INTERACTIONS FOR PRIVACY PROTECTION

Presentata da: Gaia Gambarelli

Coordinatore Dottorato

Marco Antonio Bazzocchi

Supervisore

Aldo Gangemi

Co-supervisori

Marco Lunghini

Rocco Tripodi

Esame finale anno 2023

Abstract

In recent years, there has been exponential growth in using virtual spaces, including dialogue systems, that handle personal information. The concept of personal privacy in the literature is discussed and controversial, whereas, in the technological field, it directly influences the degree of reliability perceived in the information system (privacy 'as trust').

This work aims to protect the right to privacy on personal data (GDPR, 2018) and avoid the loss of sensitive content by exploring sensitive information detection (SID) task. It is grounded on the following research questions: (RQ1) What does sensitive data mean? How to define a personal sensitive information domain? (RQ2) How to create a state-of-the-art model for SID? (RQ3) How to evaluate the model?

RQ1 theoretically investigates the concepts of privacy and the ontological state-of-the-art representation of personal information. The Data Privacy Vocabulary (DPV) is the taxonomic resource taken as an authoritative reference for the definition of the knowledge domain. Concerning RQ2, we investigate two approaches to classify sensitive data: the first - bottom-up - explores automatic learning methods based on transformer networks, the second - top-down - proposes logical-symbolic methods with the construction of PRIVAFRAME, a knowledge graph of compositional frames representing personal data categories. Both approaches are tested. For the evaluation - RQ3 – we create SPeDaC, a sentence-level labeled resource. This can be used as a benchmark or training in the SID task, filling the gap of a shared resource in this field.

If the approach based on artificial neural networks confirms the validity of the direction adopted in the most recent studies on SID, the logical-symbolic approach emerges as the preferred way for the classification of fine-grained personal data categories, thanks to the semantic-grounded tailor modeling it allows. At the same time, the results highlight the strong potential of hybrid architectures in solving automatic tasks.

Keywords: sensitive personal data, sensitive information detection, privacy corpus, transformer models, knowledge-graph, hybrid models

Abstract

Negli ultimi anni abbiamo assistito a un'esponenziale crescita dell'uso di spazi virtuali, tra cui sistemi di dialogo, che gestiscono informazioni personali. Il concetto di privacy personale in letteratura risulta parecchio dibattuto e controverso, mentre in ambito tecnologico influenza direttamente il grado di affidabilità percepito nel sistema di informazione (privacy '*as trust*').

Questo lavoro mira a tutelare il diritto alla privacy sui dati personali (GDPR, 2018) ed evitare la perdita di contenuto sensibile esplorando il task di *sensitive information detection* (SID). Si basa sulle seguenti domande di ricerca: (RQ1) Cosa si intende per dato sensibile? Come definire un dominio di informazioni personali sensibili? (RQ2) Come creare un modello allo stato dell'arte per il rilevamento automatico dei dati sensibili? (RQ3) Come valutare tale modello?

La RQ1 indaga teoricamente il concetto di privacy e lo stato dell'arte sulla rappresentazione ontologica di informazione personale. Il Data Privacy Vocabulary (DPV) è la risorsa tassonomica presa come riferimento per l'organizzazione del dominio di conoscenza. Affrontiamo la RQ2 investigando due approcci di classificazione di dato sensibile: il primo – *bottom-up* - riguarda metodi di apprendimento automatico basati su reti transformer, il secondo – *top-down* – si basa su metodi logico-simbolici e sulla costruzione di PRIVAFRAME, un knowledge graph di frame composizionali che rappresentano le categorie di dato personale. Entrambi gli approcci sono stati testati. Per la valutazione – RQ3 - è stato costruito SPeDaC, un corpus etichettato a livello di frase. La risorsa può essere utilizzata come benchmark o training nel SID task, colmando la lacuna di una risorsa condivisa evidenziata in letteratura.

Se il primo approccio basato su reti neurali artificiali conferma la validità della direzione adottata nei più recenti studi sul task, l'approccio logico-simbolico emerge come strada prediletta per la classificazione di categorie di dato personale fine-grained, grazie alla modellazione sartoriale *semantic-grounded.* I risultati del lavoro evidenziano al contempo le forti potenzialità delle architetture ibride nella risoluzione di task automatici.

Keywords: dati sensibili personali, identificazione dell'informazione sensibile, privacy corpus, modelli transformer, grafi di conoscenza, modelli ibridi

Acknowledgments

Vorrei ringraziare sentitamente chi mi ha accompagnato lungo il prezioso percorso di dottorato, tanto entusiasmante quanto sfidante, di cui questa tesi non è che il capitolo finale.

Prima di tutto, grazie al mio supervisore Aldo Gangemi. Ha saputo guidarmi con professionalità e acutezza, ma anche profonda umanità e gentilezza. Ha costantemente ampliato i miei orizzonti di ricerca, arricchendo i momenti di confronto con stimoli sempre nuovi. Farò senz'altro tesoro dell'esperienza e dell'entusiasmo che è capace di trasmettere con totale generosità.

Grazie a Rocco Tripodi, per essersi interessato un anno e mezzo fa al mio lavoro e avere prontamente messo a disposizione del progetto la sua preziosa e solida conoscenza, nonché il suo tempo. Il supporto e la collaborazione sono stati fondamentali sia per la crescita del progetto che per la mia crescita personale. Grazie inoltre dei preziosi e utili consigli su come redarre un paper scientifico.

Grazie a tutto il STLab, un gruppo di straordinari ricercatori che fanno del dialogo e della condivisione costante un'indiscutibile punto di forza. Le iniziative del gruppo alle quali ho partecipato hanno sempre suscitato in me vivo interesse, e sono state spesso occasione di approfondimento, messa in discussione e miglioramento. Il gruppo mi ha inoltre stimolata a pianificare l'avanzamento del lavoro e a credere nei risultati ottenuti valorizzandoli. Grazie in particolare a Valentina Presutti, che lo anima e dirige con passione e rigore. A Giorgia Lodi e Alessandro Russo della disponibilità e gentilezza che mi hanno sempre dimostrato. A Stefano De Giorgis, Arianna Graciotti, Anna Raimo e Bruno Sartini per avermi aiutata nella valutazione del corpus.

Un ringraziamento speciale va a tutta Ellysse srl., per avermi dato l'opportunità di seguire questo programma di dottorato e per la fiducia riposta in me fin dall'inizio. Ellysse è una realtà aziendale dalla rara apertura mentale e propensione all'innovazione. È capace di dare attenzione e valore alle persone che ne fanno parte. Vorrei riservare un ringraziamento particolare a Marco Lunghini, che per primo ha creduto in questo percorso. A lui e a Francesco Buzzoni, grazie per avermi seguita e supportata con grande professionalità nella mia nascita e crescita lavorativa, dei loro stimoli e dell'entusiasmo che dimostrano in ciò che fanno. Un grazie particolare a Giulia Montanari, alla sua competenza e dolcezza. Grazie a Mario Caruso e alla sua fantastica squadra della collaborazione.

Grazie, infine ma enormemente, a mio fratello Edoardo, ai miei genitori, Lino e Simona, e a Elisa, Sara e Arianna per desiderare sempre, così tanto, nient'altro che la mia felicità.

Grazie!
Gaia

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement and Importance

In recent years we have seen the exponential growth of applications, including dialogue systems that handle sensitive personal information [94, 2]. Identifiable individuals can reveal explicitly or implicitly inferable personal information from the texts they write and from the information they daily share online (in blogs, public pages, social media, etc.). The contexts in which personal information can be expressed concern 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 takes on importance.

In 2018, legal regulations have been introduced in the European Union (EU): with the General Data Protection Regulation (GDPR) [190] the right to privacy regarding personal data (PD) is claimed. The European regulation for the protection of PD requires the data controller to adopt adequate technical and organizational measures in order to protect data from unlawful processing. Article 25 in particular concerns '*Data protection by design and protection by default*'[1] and introduces two fundamental concepts. By privacy by design [22] we mean a principle of incorporation of privacy that must be adopted immediately in the processes by the companies and bodies responsible for the projects released. It makes use of 7 principles concerning: (i) the prevention of problems and risks related to the protection of privacy; (ii) privacy as a default setting in user requests; (iii) the use of pseudo-anonymization or data minimization techniques; (iv) maximum functionality; (v) safety throughout the product cycle; (vi) visibility and transparency of the processing; (vii) centrality of the user. Privacy by default (protection by default) intends to guarantee data processing only to the necessary and sufficient extent for the intended purposes and the strictly necessary period for these purposes.

There are cases where robust security policies, such as obfuscation or anonymization of the identifiable person, can be applied, and on the other hand, cases in which this information may be necessary, visible, or obtainable from 3rd parties. In the latter cases, it is important to design tools capable of revealing only the information strictly necessary for the objective, obscuring information with sensitive content which is not. In online conversations and unstructured text, for example, the loss of privacy can be very high [42] and the average cost of a data breach increases over years [189]. Large-scale sensitive information leakage incidents have been reported

[1] https://www.privacy-regulation.eu/it/25.htm (last access January 18, 2023)

in recent years e.g., the Cambridge Analytica incident which involved more than 50 million user information in 2018[2]. 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. To prevent the risk, measures can be adopted, however not trivial, if we think about how in most cases sensitive personal information can be transmitted today. It is estimated that 80% of the data currently disseminated on the Web is of an unstructured type [42, 4] i.e., data not present in a relation database, which can be presented in an irregular and contextual form. Processes of automatic identification of such information alleviate the demanding work of human control, even if they are very challenging automatic tasks.

Sensitive Information Detection (SID) is a task that constitutes a subpart of Data Leak Detection (DLD) dealing with the automatic identification of sensitive information. The task generally contributes to improving Data Loss Prevention (DLP) systems commonly designed by industries to help businesses avoid data breaches. It presents a way to train, classify and perform the classification of sensitive text [68].

As we will see, the SID literature presents very different works: studies on organization information [54, 121], and works on personal information conducted with a non-contextual approach [87, 156] or in a contextual way [64, 57]. They also differ concerning the type of personal information investigated: basic personal information [35, 64], personal health information (PHI) [54], ethnic, and political opinion information [57]. We will discuss this in detail in the section 2.3 dedicated to the state-of-the-art.

In addition to the scientific literature, since it is a highly applicative problem, we have looked at the tools currently released by IT companies such as Microsoft [193], IBM [192], and Google [191]. The tools offer services first for the automatic identification of personally identifiable information (PII): basic entities from which the subject can be uniquely identifiable (email, credit card number, social security number, etc.). However, there are cases in which the user is revealed and it is necessary to hide specific personal information, or where the aggregation of more complex personal information could lead to the identification of the individual. The aforementioned services propose the creation of categories of sensitive data that can be customized by identifying regular expressions or keywords. But how to automatically identify complex personal information? And how to disambiguate, with the same linguistic elements, sensitive contexts from neutral contexts?

Let's take some fictitious examples to better frame the problems that will be investigated:

My email address is alicia.example@gmail.com (a)
I'm the head of the oncology department at my city hospital. (b)
I live in Sassuolo, in the province of Modena. (c)
I live for football: I have always supported Sassuolo. (d)
How nice it would be to live in Sassuolo! (e)

Sentence (a) contains a PII identifiable through regular expressions. Sentence (b) presents information relating to the individual's profession (type of profession and workplace); sentence (c) refers to the location where the individual lives. Both sentence (b) and sentence (c) cannot be expressed through regular expressions. Identification could be aided by recurring linguistic patterns. Furthermore, if sentence (b) and sentence (c) were to occur in the same environment,

[2] https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (last access January 18, 2023)

pronounced by the same individual, the latter could be easily identifiable even if its identity was not explicitly revealed. Then there is another observation to be taken into account: sentences (d) and (e) have the same keywords as the sentence (c) (to live, City_name). However, the information that the context (the sentence) gives us makes only the first (c) sensitive, in which the verb 'to live' is not metaphorical as in (d) and the information is presented on a level of reality concerning the hypothetical and desire plane of the sentence (e).

In this project conducted on the English language, we, first of all, focused our attention on a well-defined domain of inquiry: informal online language, which we often use in chat interactions with other humans or with conversational agents in a virtual environment, whether it is synchronous or asynchronous. The type of information investigated is complex personal information and we first tried to define which categories can represent personal information, using a theoretical-conceptual analysis and authoritative taxonomic resources already present in the literature.

Therefore, the problem has been approached with a context-aware approach and the contextual level studied is the sentence. Starting from the examples cited above, we have moved our work in a twofold direction: on the one hand, the search for methods and techniques that classify a sentence as a sentence with or without sensitive content; on the other hand, the identification and automatic recognition of the type of sensitive personal information explicitly or implicitly present.

State-of-the-art work on SID has recently highlighted the validity of neural network based techniques [167, 64, 57, 121]. Language Models (LMs) present a very strong sensitivity to context (section 3.4). We have tried to solve our problem first by applying this approach and confirming its validity in the domain we investigated, and second by exploring and evaluating original approaches based on rules and ontological models. This kind of approach, which is characterized by top-down modeling, has the particular advantage of not requiring training data. The model also has the advantage of being tailored: it offers customizable sensitivity labeling according to the needs of obfuscation and protection.

Indeed, a further problem that we had to address can be introduced; it has already been highlighted in the literature [120] and will be discussed extensively in the dissertation: the lack of a benchmark and available labeled sensitive data resources. This is especially true, as it could seem obvious, for the personal information domain. This work makes its contribution by releasing the resources built for the resolution of the task on the categories of PD. The datasets have been built by collecting real texts of informal language downloaded from Web pages of different types, and do not present any kind of reference to identifiable individuals. The annotations are made both to discriminate sensitive vs. non-sensitive sentences, and to classify the different types of sensitive categories; the resource has been evaluated on the proposed models. Datasets are available to be shared; they can therefore be used as a training set for training new models and as an evaluation benchmark for new approaches.

1.2 Research Questions

The following research aims essentially to contribute to the state-of-the-art in automatic PD detection from text, investigating hybrid models and techniques for a context-aware approach. The work addresses the following research questions:

RQ1. What does sensitive data mean? How to define personal sensitive information domain?

RQ2. How to create a state-of-the-art model for automatic sensitive data detection?

RQ3. How to evaluate the model?

These are the fundamental questions that open to the reflections and contributions of the work. In the final section 5.2.1, they will be resumed and discussed.

1.3 Thesis Structure

The premises just discussed are followed by the survey and the discussion on the theoretical implications, the state-of-the-art of the addressed problem, and by the presentation of the research contributions.

The dissertation is structured as described:

Chapter 2. Background - The aims of this chapter are the following:

- To give a theoretical perspective of the complex concept of privacy. Its meaning is explored by adopting two main perspectives: the theoretical [2.1.1] and the legal [2.1.2] ones (with an in-depth analysis dedicated to the GDPR [2.1.2.1]). When we talk about privacy we talk about an open-texture problem often very difficult to clearly define.

- To reflect and define to what sensitive personal data (SPD) means [2.2]. This section helps us also to understand what characterizes sensitive personal information, to better design the model of automatic identification.

- To outline a complete state-of-the-art of the SID task [2.3], discussing the related work in relation to the different approaches [2.3.1] which can be considered; we present the works based on Natural Language Processing (NLP) and Machine Learning (ML) approaches [2.3.1.1], the inference rule-based and ontological approaches [2.3.1.2], the most recent works based on neural networks and deep learning [2.3.1.3], and finally the application of hybrid approaches, which combine statistical with logical and symbolic approaches [2.3.2].

- To introduce an overview of the privacy ontologies proposed for privacy protection [2.4, 2.4.1]. In particular, attention is given to the authoritative resource used as a reference for the analysis and selection of PD categories to investigate and to model in our work, the Data Privacy Vocabulary [2.4.2]: it is presented with an introduction [2.4.2.1], a description of its structure [2.4.2.2], the PD categories it involves [2.4.2.3], and finally the contribution to our work [2.4.2.4].

- To outline a complete state-of-the-art of the privacy corpora and to highlight at the same time the lack of a shared resource for PD [2.5]. The corpora most often used in literature are the Enron corpus [2.5.1] and the Monstanto dataset [2.5.2], but they are labeled for organizational sensitive information. Other corpora explored in literature are described [2.5.3], but they are not available or they do not present labels of our interest. For that reason, we dedicate a section to discuss the limits [2.5.4] and to introduce our contribution to fill the gap.

Chapter 3. Methodology - The aims of this chapter are the following:

- To present the methodology adopted to address the task, describing a twofold approach: a knowledge-driven top-down and a data-driven bottom-up approach [3.1].
- To highlight the interest and the advantages to implement a hybrid approach, considering and combining deep learning and symbolic methods [3.2]; in this section, we also introduce the structures of the models proposed.
- To detail the frame-based approach starting from its theoretical bases and applying it to our problem [3.3]. In this section little space is given to the introduction of the three main semantic resources used in our model: FrameNet [3.3.1], WordNet [3.3.2], and Framester [3.3.3].
- To deepen the transformer models approach starting from its theoretical bases and applying it to our problem [3.4].

Chapter 4. Contributions - This chapter is constituted of three different types of original contributions. The aims of this chapter are the following:

- To present a new resource for sensitive personal categories proposed to address our problem, as an evaluation dataset for our models, and to fill the lack of available resources and benchmark in literature [4.1]. The in-depth aspects concern the ethical disclosure [4.1.1], the analysis of which kind of personal categories from Data Privacy Vocabulary (DPV) are to take into consideration [4.1.2], the structure of the resource [4.1.3], the description of the methodology followed to collect and label data [4.1.4], and finally a discussion of relevance and limits of the contribution [4.1.5].
- To present the transformer-based model for the sensitive vs. non-sensitive and macro personal categories classification [4.2]. The section dedicated to methodology presents how the transformer model can address the task, putting it into the hybrid loop [4.2.1], and is followed by the description of the configuration of the model based on RoBERTa. The model has been evaluated on the aforementioned resource: the experimental process [4.2.3], the results [4.2.4], and the analysis of the results [4.2.4.1] are described. Even this section concludes by presenting the relevance and limits of the approach and considering the achieved results [4.2.5].
- To present the model knowledge-graph-based for the fine-grained identification of SPD. It is introduced by a description of the structure [4.3.1], followed by a methodology section [4.3.2]. Its evaluation of the created resource is presented in the experimental process section [4.3.3], showing results [4.3.4] and analysis [4.3.4.1]. As for the other contributions, the last section is dedicated to relevance and limits [4.3.5].

Chapter 5. Discussion and Conclusion - This chapter presents a discussion on the studies conducted outlining some emerged reflections [5.1]. A section of conclusions [5] points out the results obtained with respect to the state-of-the-art and the research questions [5.2.1]. Finally, future works are presented. These are divided into (i) ideas to improve the models; (ii) horizons to be explored that emerged in the discussion useful to broaden the contributions or deepen certain aspects of them; (iii) future works contributing to the SID task.

Chapter 2

Background

In this chapter, we intend to analyze in detail the background concerning the automatic identification of sensitive data and privacy protection tasks. In section 2.1 we will explore the definitions that have been used to describe such a complex, broad-boundary, and controversial concept as that of privacy. We will see how this theoretical vagueness can be partly averted when the legal perspective is taken into account; even if the legal domain itself is considered in the literature and by its nature an open-texture problem. The theoretical perspective is useful for us to realize the complexity and nuances of the problem, which are undoubtedly reflected in the resolution of more pragmatic automatic identification tasks.

In particular, we will start from the recent normative on the protection of PD, the GDPR, to define our investigation domain in section 2.2 and analyze in depth the personal data categories (PDCs).

In section 2.3, an overview of the works in the literature on the task of sensitive information identification is presented, considering the many perspectives and the different approaches adopted and presenting the works logically divided by techniques/models used. The review is very important because it establishes the foundations of our work: we draw inspiration from previous studies by identifying the techniques that have reported the greatest successes, the limits that can be shared by more than one work, and the state-of-the-art achieved so far.

In section 2.4, a survey of the ontologies that have been developed for privacy in the broad sense will be presented (often these ontologies concern the interpretation of laws and regulations). In particular, the Data Privacy Vocabulary is described in detail; it is the richest and most complete state-of-the-art ontology that conceptually represents PD and the rules regulating its use. This ontology will be very relevant to our work.

Finally, in section 2.5 we address one of the biggest problems in literature: the lack of labeled corpora and a shared benchmark. We describe the resources developed and adopted so far for the evaluation of the proposed models, introducing our contribution in this sense.

2.1 The concept of privacy

What do we mean by privacy? What is its definition? What theoretical concepts emerge and are important for its practical treatment? This brief and purely theoretical chapter, which can be considered a real premise, aims to illustrate how the concept of privacy can be understood, what its boundaries are, and whether it is possible to give it a univocal definition.

2.1.1 Theoretical perspective: privacy definitions

'Privacy is a concept in disarray' states Solove [157]. Philosophers, lawyers, jurists, and theorists have spent their time trying to define the concept of privacy. In this sense, literature is dense with definitions. The academic and jurist Miller [111] defines it as *'vague and evanescent'*; for the academic Lillian BeVier it is instead:

> *A chameleon-like word used denotatively to designate a wide range of wildly disparate interests - from confidentiality of personal information to reproductive autonomy - and connotatively to generate goodwill on behalf of whatever interest is being asserted in its name.* [15]

The definition of privacy can still be said as an open question with many unanswered questions [103].

Two interesting privacy frameworks are given to us by Westin and the aforementioned Solove. Alan Westin, one of the fathers of modern privacy law, gives us the definition in his most famous book *'Privacy and Freedom'* [180], which is the following: *'the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others'*. He presents a hierarchical framework, starting from the distinction between political, socio-cultural, organizational, and personal levels. The concept of privacy is built and concretized in each of these levels and it is therefore very clear that a base for defining privacy concerns the context in which it is located. Solove also proposes a framework for understanding privacy contextually, arguing that privacy theories are all either too limited - while claiming to be exhaustive - or too vague to be effective. He proposes a model based on Wittgenstein's theory of similarities:

> *I suggest abandoning the traditional way of conceptualizing privacy and instead understanding it with Ludwig Wittgenstein's notion of 'family resemblances'. Wittgenstein suggests that certain concepts might not have a single common characteristic; rather they draw from a common pool of similar elements. Privacy, therefore, consists of many different yet related things [. . .] I argue that privacy should be conceptualized from the bottom up rather than the top down, from particular contexts rather than in the abstract.* [157]

Solove, therefore, defines privacy in a *'pluralistic manner'*, and as a set of protections that concern very different problems, albeit well correlated (processing, invasion, treatments, etc.). It is evident that in each case, the concept of context is central. The definition of privacy is bound to change concerning the application context over time.

Although the debate on the definition of privacy is not so recent, De Hert and Gutwirth argue that it is nevertheless a relatively new concept in the development of contemporary law [33]. From a purely legal point of view, we often do not speak of *privacy* but of more specific and pragmatic terms, such as *data protection* (DP), a concept closely linked to individual protection. DP is a means of protecting the value and interest of identity, security, and freedom of information [5]. The protection of PD concerns the processing of all information relating to an individual and is, therefore, more defined than the abstract concept of privacy, which can sometimes also imply purely personal choices. Privacy, following this reflection, could therefore be considered a moral right - more than anything else - that preserves the autonomy of the individual [159].

Surely the urgency of privacy and protection is a topic on which attention has grown strongly with the advent and pervasiveness of technological tools and applications, such as social networks and more generally virtual spaces in which information is exchanged. For example, the spread of dialogue systems based on artificial intelligence, chatbots and voicebots, adopted by both companies and public administrations [160, 6], which allow H-R interaction in natural language, is increasingly established. When it comes to Artificial Intelligence (AI), privacy is now a central aspect and in close connection with the concept of trustworthiness. AI-based technologies in particular should respond to the concept of reliability. This reliability is also influenced by the transparency and attention that technology places on the management of data and information deemed personal or confidential. If it is not possible to arrive at a convincing definition of privacy in its broadest sense, we can try to do so by limiting it to specific new technologies and AI domains. In the technological field, and in particular, in the AI technological field, a new definition of privacy can be imagined: the concern for privacy as something strictly related to the value and protection of the identity, security, and freedom of the person, which assumes at the same time the character of privacy as trust, influencing the degree of reliability perceived in the information systems.

We can even find some studies that try to measure the privacy perception of people through linguistic techniques. They found their work on the prototype theory evolved from Wittgenstein's family resemblance, which proposes to define concepts through prototypes that represent the average member of a concept [181]. Vasalou et al. [173] constructed a model to automatically detect relevant discourse about privacy: they collected a list of features over privacy concepts and centrality rates of features regarding the concepts by participants. They define totally 82 privacy features that provide a very inclusive and qualitative conceptualization of privacy. The model has been evolved in a follow-up work by the authors [174].

Beyond morals or shared perceptions, therefore, a new interdependence is established in this sense; a necessary coexistence and dialogue between law and technology. It can be defined as a mutual benefit relationship: law can describe and regulate the protection of privacy in the virtual and technological world; technology can help privacy, in the strictest sense of data protection, on multiple levels (processing, obfuscation, sanitization, etc.).

In the next paragraph, the legal systems that outline privacy as a right will be deepened. However, even if the thesis has a more experimental and pragmatic structure, a theoretical framework in a broader sense helps us to understand the complexity of the problem we are about to face. This complexity is necessarily reflected in the many different perspectives that related works on privacy protection adopt (data identification, data processing, privacy policies treatment, protection of textual, biometric data, etc.). Even the circumscribed field of textual data protection is characterized by substantial contextual differences, which, to return to Westin, first concern the level to be investigated (personal, organizational, or political), making it difficult to compare the works of the same task with each other; but above all, these reflections force us to start from a fundamental question: what then can we consider sensitive?

2.1.2 Legal perspective: an open-texture problem

Even in the legal field, the definition of privacy does not seem obvious at all. To emphasize the gray areas that have already appeared during the search for a definition of privacy in the previous paragraph, and to normalize and extend this vagueness to the more generic legal domain, we introduce the concept of the open-texture problem (OT).

To introduce the concept of OT or conceptual porosity is Waismann [177] who argues that *'most of our empirical concepts are not delimited in all possible directions'* and that the definitions are, instead, *'always correctable or amendable'*. For example, let's take a plant that reaches a much larger size than all the other plants of its species. For this extreme case, should we redefine the concept of plant or rather not define it as such? According to the philosopher, simply, an empirical definition can never be exhaustive i.e., capable of providing necessary and sufficient elements for it to be defined as complete. The new or borderline possibilities would always leave the concept open and its applicability or not.

Open Texture
⇑
Incomplete definition
⇑
Incomplete description
⇑
Open texture
[77]

The concept of OT opens the debate in various contexts, including the philosophy of legal language and the philosophy of information and artificial intelligence.

Two types of OT can be distinguished. The first is the one theorized by Waismann himself, while the second is theorized by Shapiro [154] and is similar to the concept of porosity:

1. **(OT1)**. A concept or term shows an OT if objects that do not fall under the standard application domain are possible for which there is no factual data as to whether they fall under the concept or not.

2. **(OT2)**. A concept or term shows an open plot if there are instances for which even a competent and rational agent can acceptably state that the concept applies or does not apply.

In the first case, the concept of opening is introduced, while in the second that of porosity. And if in OT1 the concepts can be expanded if applied to new domains, in OT2 the porous concepts can be better defined by solving the borderline cases within their domain of application.

Waismann and Shapiro have conflicting views on solving concepts at OT. Waismann contrasts the concept of vagueness and that of openness, arguing that if the former can be resolved by refining and solving all borderline cases, this does not mean that the risk of OT is averted, as each new case could make the definition uncertain again. For Shapiro, however, vague concepts would always be bound by context. The only way out of the impasse would be to consider new or borderline concepts not as such, but as elements of a new whole.

The OT in the legal domain is a theory formulated by Hart [67], who describes all legal norms as *'characterized by a penumbra of uncertainty'*. Reflecting on judicial decisions, he wonders about the risks of formalism:

> *But just how in being a formalist does a judge make an excessive use of logic? It is clear that the essence of his error is to give some general term an interpretation which is blind to social values and consequences (or which is in some other way*

> *stupid or perhaps merely disliked by critics). But logic does not prescribe interpretation of terms; it dictates neither the stupid nor intelligent interpretation of any expression. Logic only tells you hypothetically that if you give a certain term a certain interpretation then a certain conclusion follows. Logic is silent on how to classify particulars-and this is the heart of a judicial decision. So this reference to logic and to logical extremes is a misnomer for something else, which must be this.* [67]

OT is also an established problem in legal information systems, referring to all systems that are unable to deal with heterogeneous problems. In this context, questions were raised about how rule-based, statistical, case-based, or ontological systems can deal with this issue [13]. In rule-based systems the concept is applicable if sufficient conditions are satisfied, otherwise, it is not [150]. An exclusive system acts within the core of certainty, tacitly excluding the penumbra cases. A non-exclusive system, on the other hand, can consider borderline cases by proposing both the conditions of applicability and those of inapplicability by leaving the resolution to the user or automating the decision with a series of metarules and quantitative weights that determine the various reasons of conflict [65]. The case-based approach is based on the consideration that a case can be solved concerning similar cases that occurred previously, using statistical-quantitative and qualitative measures (meta-principles that prefer certain cases, such as the most recent, or attribute weights based on the relationships they present). The statistical approach will naturally be quantitative. The last approach is the ontological one. We will introduce the main concepts of ontology in a later dedicated paragraph (section 2.4). However, here we refer to two ontologies developing for the legal domain: the functional legal ontology of Valente and Breuker [20] and the frame-based ontology of Van Kralingen and Visser [176].

Valente and Breuker distinguish six types of legal knowledge categories:

- The normative knowledge defines a standard of social behavior;
- The world knowledge describes the world that is been regulated;
- The responsibility knowledge assigns or restricts the responsibility of an agent for its behavior;
- The reactive knowledge concerns the consequences of the violation of a norm;
- The meta-legal knowledge refers to legal knowledge in a meta sense;
- The creative knowledge is the legal knowledge that allows the creation of previously nonexistent legal entities.

On the other hand, Visser and Bench-Capon define three concepts in their ontology, each described through the representation of frames:

- Norms: general rules and standard principles of conduct to which the subjects must comply;
- Acts: the dynamic aspects that can change the state of the world. There are various discriminants to classify the aspects: a. events (instantaneous change between two states) and processes (duration), b. institutional and physical acts;
- Descriptions of concepts: the meaning of concepts that can be found in the domain.

Therefore, only the statistical approach tries to solve the open-texture problem in an exclusively quantitative way. Also in the paper by Mullingan et al. [116], the authors reflect on the conceptualization of privacy as an open-texture problem. In particular, they report the Gallie framework consisting of seven analytical criteria that define privacy as a contested concept:

- Appraisiveness: concepts provide value judgments;
- Internal complexity: it denotes fully assessed entities characterized by internal and multidimensional complexity;
- Diverse describability: concepts can be described differently by those who discuss them;
- Openness: concepts change according to time and circumstances;
- Reciprocal recognition: concepts can be used toward others both aggressively and defensively;
- Exemplars: concepts must derive from authoritative examples and be acknowledged by all disputants;
- Progressive competition: contesting concepts contribute to their continuous change.

Based on this, they propose an *ad hoc* analytical method to deal with privacy and its controversial aspect. The framework proposes five dimensions to analyze privacy:

- Dimension of theory: object (what's privacy for? E.g., dignity, control over personal information); justification (why should this be private? E.g., individual liberty, social welfare); contrast concept (what's not private? E.g., public, open, transparent); exemplar (what's an example? E.g., identity theft, intrusive surveillance).
- Dimension of protection: target (what's privacy about? Privacy of what? E.g., personal information, private space); subject (whose privacy is at stake? E.g., myself, my child, social groups).
- Dimension of harm: action (what act violated privacy? E.g., collection, processing, dissemination, and invasion); offender (who violated privacy? E.g., government, business entity); from whom (who is privacy-protecting against? E.g., government, everyone).
- Dimension of provision: mechanism (how is privacy provided? E.g., legal regulations, social norms); provider (who is supposed to provide privacy? E.g., government, business entity).
- Dimension of scope: social (where is privacy found? E.g., hospital, university, state); temporal scale (how long is privacy required? E.g., permanent, variable expiration); quantitative scope (how widely does privacy apply? E.g., universally as a strict rule, casuistically as per-case).

All the factors listed in the framework just mentioned help to delineate the dark areas that make privacy an open-texture problem. The task of identifying privacy starting from text, whether it takes into account the context or not, presents as an aspect of substantial importance the target, belonging to the protection dimension, for which we must first determine the domain of sensitive data we are investigating; the subject, of the same size, already discriminates contextual approaches from those that are not. All other aspects are not directly involved in the interest of solving our task, although they can equally contribute to the introduction or not of cases to be taken into consideration.

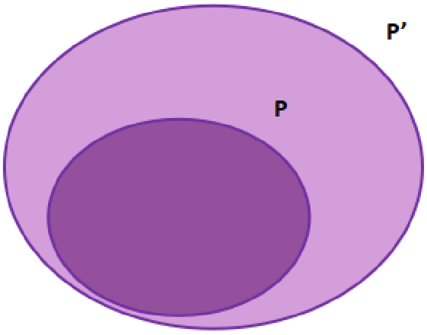


Figure 2.1: Privacy identification from text in OT

The OT problem of textual identification of privacy can be represented as shown in Fig. 2.1. P can be considered as a zone of certainty, in which the protection dimension is well defined: the target can be defined as t and the subject as s. *P'* represents the penumbra zone, and could at least coincide with the following cases:

- t is well specified, but s is not;
- s is well specified, but t is not.

2.1.2.1 General Data Protection Regulator (GDPR)

From a legal point of view, the concept of privacy finds more precise definitions up to exceeding the very definition of '*privacy*' in the GDPR. Privacy begins to be considered in the legal systems of the EU since 1981, in response to technological advances, with the Convention for the 'Protection of individuals with respect to the automated processing of PD'. This convention aims to regulate the processing and free flow of PD within the EU.

Convention 108 already defines what a PD is, that is, '*a data relating to an identified or identifiable natural person*'. It is related to the right to freedom i.e., the protection from external control (private or state). These concepts are translated into a directive in 1995 (Directive 95/46/EC) of the Council of Europe. This directive was revised in 2002 in the light of the increasingly considerable implications given by the technological pervasiveness, the E-Privacy Directive, on the processing of PD in relation to electronic communication and then by the Directive 2009/136/EC, which additionally regulates the a priori consent of the user before the installation of cookies on the electronic device[1].

When the Treaty of Lisbon was signed in 2007, which makes the Charter of Nice (2001) legally valid, Article 8 of the latter on the protection of PD enters into force. This is where we start talking about the processing of PD. The GDPR never mentions the extended word '*privacy*'. It comes out in 2016 and enters into force on May 25, 2018. There are two premises of the GDPR (Recital) underlined here:

[1] Directives are binding on states in aims, but not in means; that is, they leave the individual countries a very wide space of interpretation and the possibility of adopting national regulations considering the indications of the directive. Regulations, on the other hand, like the GDPR, are binding throughout the EU.

> *The processing of personal data should be at the service of man. [...]* (GDPR, Recital, 4)

> *The rapidity of technological evolution and globalization bring new challenges for the protection of personal data. The scope of sharing and collecting personal data has increased significantly. Current technology allows both private companies and public authorities to use personal data, as never before, in the conduct of their business. Always more often, individuals make personal information about them available to the public worldwide. Technology has transformed the economy and social relations and should facilitate even more the free flow of personal data within the Union and their transfer to third countries and international organizations, while ensuring a high level of protection for individuals. personal data.* (GDPR, Recital, 6)

The GDPR, despite being a regulation, leaves some aspects to national legislation in the context of the processing of sensitive data, the discipline of the Guarantor Authority, the sanctions, and the laws of subjects other than Parliament (such as professional categories). The GDPR differentiates, in particular, two types of data:

- **Particular data** i.e., data that must not be processed without the explicit consent of the interested party and specific needs. Particular data include racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic and biometric data intended to uniquely identify a natural person, data relating to health, sexual life, or sexual orientation;
- **Personal data**, which with technological evolution is something more complex. By PD, the GDPR means:

  > *[...] any information concerning an identified or identifiable natural person (interested party); the natural person is considered identifiable who can be identified, directly or indirectly, with particular reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more characteristic elements of his physical identity, physiological, genetic, psychic, economic, cultural or social.* (GDPR, Article 4.1)

  These data must always be treated in accordance with the GDPR.

Both particular data and PD are sensitive. Other key terms of the GDPR are:

- **Controller**, Article 4-7: the entity or organization that determines the purpose and means of the processing of PD;
- **Processor**, Article 4-8: the entity which processes PD on specified instructions of the controller;
- **Data Protection Officer (DPO)**, Article 37: an individual appointed by a controller or processor to oversee compliance and processing of PD, monitor internal processors, and collaborate with supervisory authorities as required;
- **Regulatory or Supervisory Authority or Data Protection Commission**: a governmental organization with the responsibility to evaluate and enforce GDPR compliance. These bodies are established by national or federal governments and have jurisdiction over their appointed regions.

The privacy policy must be guaranteed by the data controllers and is aimed at informing the data subjects of the identity of the data controller(s), the purpose of the treatment, legal basis, transfer of data and categories of recipients, data retention period and processing rights. Particular PD then contain additional obligations to be considered in the processing.

The resource that we will use as a reference for conducting our work does not exclusively follow the GDPR legislation. However, this legal rule allows us to highlight the following aspects:

- Attention to a less evanescent domain and dependent on socio-cultural factors such as the one discussed so far on *'privacy'*;
- The importance of the protection of personal information is recognized and defined at the legislative level;
- Importance and impact of the problem in relation to increasingly pervasive technological advances.

From a regulatory point of view, therefore, we can rely on a much more crystalline framework and it is from this that we start defining our target domain (t) e.g., SPD, in the next section 2.2.

To sum up:

- The concept of privacy in a broad sense is controversial and *'evanescent'* and the theorists - while not arriving at a precise and exhaustive definition - have proposed frameworks of interpretation in which the contextual aspect is central;
- Privacy - in a broader sense - can be considered more than anything else a moral right that preserves the autonomy of the individual;
- Privacy when placed in the legal context is a relatively new concept;
- With the advent of technology, privacy has become a central theme, outlining new perspectives such as '*privacy as trust*';
- The dialogue between law and technology, in which privacy can be conceived in more pragmatic terms (e.g., Data Protection), becomes urgent;
- The complex theoretical foundations are also reflected in scientific works and on much more specific tasks for privacy protection. What we mean by sensitive data is the first problem to address;
- Concepts with complex and controversial definitions can be discussed from the perspective of OT problem;
- The problem of OT is recognized in the legal domain;
- The relative concepts of OT always present a certainty of a penumbra zone;
- The problem of privacy can be considered OT;
- The dimensions that influence the areas of certainty and those of penumbra have been defined in the more limited problem of textual identification of sensitive data: the target t and the subject s;

- The most recent legislation that recognizes the protection of SPD from a legal point of view is the GDPR;
- In the GDPR, the concept of '*privacy*' is never contemplated, while technical definitions are provided on what is meant by PD and on agents and actions involved in the data processing.

2.2 Sensitive Personal Data

In this paragraph, we try to define our investigation target: the domain of SPD.

The DPV (described in section 2.4.2) provides the following definitions (see also Fig. 2.2):

- **Data**: a broad concept representing data or information;
- **Personal data**: data directly or indirectly associated or related to an individual (GDPR Art.4-1). It is distinguished from non-PD;
- **Anonymized Data**: PD that have been (fully and completely) anonymized so that it is no longer considered as PD; PD that has undergone a partial (incomplete) anonymization process - such that it is still considered a PD - can be called pseudo-anonymized data;
- **Collected Personal Data**: PD that have been collected from another source such as the Data Subject;
- **Derived Personal Data**: PD that have been obtained or derived from other data;
- **Generated Personal Data**: PD that have been generated or brought into existence without relation to existing data i.e., data not derived or inferred from other data;
- **Inferred Personal Data**: inferred PD are derived data generated from existing data, but which did not originally exist within it e.g., inferring demographics from browsing history;
- **Sensitive Personal Data**: PD considered '*sensitive*' in terms of privacy and/or impact, which therefore require additional considerations and/or protection. Sensitivity is a matter of context, and may be defined within legal frameworks. For GDPR, Special categories of PD are considered a subset of sensitive data. To illustrate the difference between the two, consider the situation where LOCATION DATA is collected and which is considered '*sensitive*' but not '*special*'. As a probable rule, sensitive data require additional considerations whereas special category data requires additional legal basis/justifications;
- **Special Personal Data Category**: SPD whose use requires specific legal permission or justification (GDPR, Art.9).

Taking up the conceptualization given to privacy in paragraph 2.1.2, PD can be conceptualized in terms of OT in this way: the core of certainty is made up of SPD, while the other PD rest in the penumbra zone and the sensitivity depends on the definition of context (Fig. 2.3).

Starting from this assumption, another definition useful for solving the automatic task of sensitive data identification can be introduced: the concept of context-aware sensitivity.

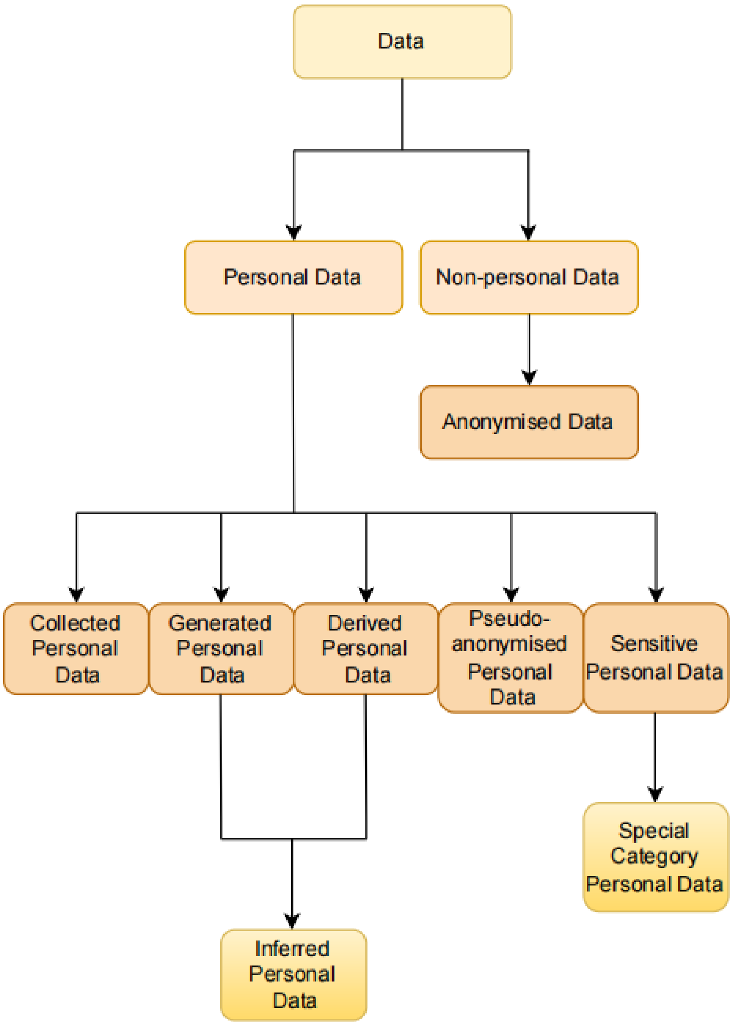


Figure 2.2: Types of PD following the hierarchical proposal of DPV

> *Two years ago I was diagnosed with breast cancer.* (a)
>
> *In the past two years, 11.7% of women have been diagnosed with breast cancer.* (b)

From a linguistic point of view, the terms used are the same in (a) and (b), but only (a) can be considered a sentence with potentially sensitive content.

Therefore, the following aspects have to be considered:

1. **Information**: the type of information explicitly or implicitly evoked in the text and potentially sensitive or not;

2. **Subject**: the subject(s) whose sensitive information must be protected;

3. **Relationship**: the relationship between potentially sensitive information and the subjects in need of protection.

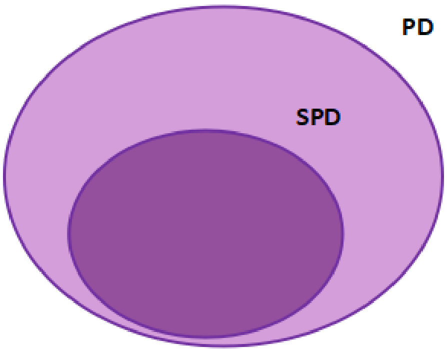


Figure 2.3: SPD and PD in OT

Based on this, assuming an investigation domain D, *I(D)* can be defined as the potentially sensitive information of that domain, with respect to all information *I*. *S(D)* are the subjects concerned with the protection of sensitive information, concerning all the subjects S elicited in the text. Given *I(D)* and *S(D)*, *I(D)* will become sensitive and not only potentially sensitive, if and only if there is an explicit or implicit relationship R for which the sensitive information *Is* refers to the subject *S(D)*.

$$D = \{I = Is \textit{ iff } (I, S) \in R\} \quad (s1)$$

However, even this definition presents areas of penumbra. Let's take these sentences as examples:

I am a criminal lawyer. (c)

I hope I can become a criminal lawyer. (d)

The function (*s1*) is true if applied both in (c) and (d); the two sentences present potentially sensitive personal information (professional occupation) and a sensitive subject directly related to it (I). Sentence (d) however expresses a desire of the subject and cannot be considered as sensitive as sentence (c) which moves into the factual reality. Borderline cases could include hypothetical situations, as well as ironic or not proven assumptions. We, therefore, introduce a variable that refers to the condition of reality r of the affirmed facts, whose truth is expressed by the value *1*:

$$D = \{I = Is \textit{ iff } [(I, S) \in R] \wedge [r = 1]\} \quad (s2)$$

Following these reflections, we could at least determine two types of context:

1. **Domain context**: the type of sensitive information to be protected;
2. **Textual context**: the relationships that potentially sensitive information interweaves with other elements of the text; in particular, its relations with potentially sensitive subjects and its conditions of reality.

As regards (1), the domain analyzed in this work concerns SPD. However, there are different types of SPD e.g., personal health information (PHI), PII, etc. Our contribution will examine personal data categories (PDCs) in a broad sense. The methodological details, concerning (1) and (2), will be described in the dedicated chapter (chapter 3).

To sum up:

- From a hierarchical point of view, we consider the broadest set of PD as potentially sensitive data, which only acquire sensitivity if certain variables are present;
- From a linguistic point of view, we can identify three fundamental aspects of sensitivity identification: the type of information, the subject, and the relationship between them (s1);
- To the three aspects mentioned above, we can add the conditions of reality of the information (s2).

2.3 Sensitive Information Detection

As previously seen, the exponential growth of applications and dialogue systems handling personal sensitive information has highlighted the important issue of data protection.

The main literature on the subject of privacy protection concerns methods of anonymization [106, 117], automatic privacy policies processing [169, 19], or studies about privacy risks and privacy leak detection [28, 184]. At the same time, shared taxonomies have been developed for the interoperability of data privacy (see section 2.4).

SID task concerns the identification of those parts of text considered sensitive in a particular context of application. To be treated and protected this information must be identified first.

In this section, we will deepen the related works and the state-of-the-art of SID task, identifying and describing its main approaches and finally introducing our contribution.

2.3.1 Approaches and state of the art

The discriminants we will refer to mainly concern the techniques used to solve the task. However, other types of differentiation can be found in related works:

1. **Domain of investigation**. A lot of work in the literature [120, 182, 68] concerns the SID task, however only some of these focus on the specific domain of personal information: basic personal information [35, 64], PHI [55], ethnic origin and political opinion information [57];
2. **Language**. SID has been investigated in several different languages. The works are mainly in English [68, 120, 64, 37, 167] and Chinese [182, 100], but also Portuguese [35], Spanish [55] and Amharic [57].
3. **Context of identification**. Sensitive data can be identified in different contextual manners. Some works consider the sensitivity of a whole document [68]; others refer to a sensitive sentence level [121, 167] or an entity level [26, 64].

However, a discriminant that will guide us concerns the non-contextual or contextual approach in order to solve the problem [120]. The studies can be divided into two macro approaches:

- **no context-aware approach**, where sensitive information does not depend on the context in which it appears; e.g., a word can be identified as sensitive regardless of the sentence in which it is present;
- **context-aware approach**, where the sensitivity of data varies according to the context. Only given the sentence, the sensitivity of a given word can be inferred.

The first no-context-aware approach includes work based on the identification of a fixed context with n-gram techniques [68, 78] or rule-based inferences to identify contextless words with sensitivity scores [26, 58]. The contextualized approach appears in the literature with the embedding technique for the recognition of a fixed context [104] or automatic paraphrasing methods using recursive neural networks [120].

In more recent years, privacy-preserving studies have mainly implemented Deep Learning (DL) methods and NLP techniques. Just to name a few, we see how Convolutional Neural Network (CNNs) [95] have been used for the sensitive detection of military and political documents in the Chinese language [182]. Bidirectional Long-Short Term Memory (Bi-LSTM) neural networks [152] have been used in a study conducted in the Chinese language on unstructured text [100] and for the identification of PD in Amharic text [57]. Transformer architectures have been used also in recent studies. A study conducted in Spanish [55] uses a BERT-based sequence labeling model to detect and anonymize sensitive data in the clinical domain. A recent study on the English language [64] proposed ExSense, a model named BERT-BiLSTM-Attention to extract sensitive information with NLP techniques from unstructured text. Timmer et al. [167] use a BERT pre-trained model to predict the Monsanto categories.

For this reason, it seemed appropriate to conduct a systematic review of the works in SID, discriminating them in relation to the methods used:

1. NLP and ML approaches
2. Inference rule-based and ontological approaches;
3. DL approaches;
4. Other hybrid approaches.

Look at the Table B.1 to have an overview of all the related work in literature and the discriminating characteristics mentioned so far.

2.3.1.1 NLP and ML approaches

Named Entity Recognition approach. Named Entity Recognition (NER) concerns the automatic identification and extraction of entity chunking in text. Entities are categories of related concepts e.g., person, locations, time expressions, quantity, etc. This task pertains to Information Extraction (IE) and it is typically faced through three methods:

- Hand-made rule-based method;
- ML-based method (the most used in literature and NLP);
- Hybrid methods

.

In a related paper [87], the authors explore the domain of PII identification. The authors issue the problem in two ways: the first one concerns the identification of PII through NER; the second one explores the relations extraction (RE) between entities through ML methods. The identified entities are person, organization, email, address, phone number, money, date, credit card number, and social insurance number. Named entities are recognized by the patterns matching technique, using regular expressions and a gazetteer i.e., a list of words that represent each entity. Instead, RE regards any pair-wise relations that may uniquely identify a particular person, e.g., personal phone number, personal email, birth date, personal income, etc. The first evaluation dataset comes from the Carnegie Mellon University; it is semi-structured and mainly constituted by documents as research papers; the second one is from the Enron Email corpus (see section 2.5). The datasets have been semi-automatic annotated. The best accuracy obtained achieves the 93.3%.

Sokolova et al. [156] the focus on the medical domain. They propose a model to detect and learn from PHI characteristics; furthermore, they found some evaluation metrics for the performance of a classification algorithm with a small sample of training data. They used a dataset from 407 Canadian documents. A PHI information is composed of a relation between PII and health information (HI):

$$PHI = IdentifyingInformation \wedge HealthInformation$$

The PII can include for example personal name, physical condition, and street address; the HI can concern disease symptoms, health care providers, behavioral state, phone numbers, or mental state. The metrics they propose concern the probability to detect PII and HI together (true detection probability, TDP) and the probability to find PII without HI (false referral probability, FRP). The detection of this sensitive information is conducted by the rule-based method NER. The NER approach has been placed in this section since in most cases the applications are based on ML; in this specific case, however, it is more appropriate to refer to the next section, dedicated to rule-based approaches (section 2.3.1.2). Some REs have been modeled in order to discover relationships between PII and PHI.

N-grams approach. The n-grams technique means a sequence of n items from a given sample of text or speech, where typically these items are words. N-gram models typically deal with natural language sentences, using statistical properties. The n-grams model can be traced to a mathematical theory [153], which assumes that if we train the model with the probabilities of all the possible next letters or words, then, given a sequence of letters or words, we can predict the probability of the next unknown ones. The theory also follows the Distributional Hypothesis [66]: '*Words (meaning) are similar if they appear in similar contexts*'. Each word depends only on the last $n-1$ words (Markov models). The model predicts x_i in terms of its probability $P(x_i|x_{i-(n-1),\dots,x_{i-1}})$. The n defines the number of words considered by the model e.g., digrams, trigrams, etc. The unigrams model assumes that the probabilities of tokens in a sequence are independent:

$$P_{uni}(t_1t_2t_3) = P(t_1)P(t_2)P(t_3)$$

Regarding the contextual approach point of view, the n-gram approach can be defined as no-context-aware. Through the n-gram approach, the word is considered in a fixed and rigid context that corresponds to the number of grams of the model.

In Hart et al. [68], the authors present algorithms of sensitive document classification considering an enterprise domain. Enterprise networks and computers present three types of data: public enterprise data (e.g., public web pages, emails to customers and other external entities, public relations blog posts); private enterprise data (e.g., internal policy manuals, legal agreements, financial records, private customer data, source code or other trade secrets); non-enterprise data (e.g., personal emails, Facebook pages, news articles, and web pages from other organizations). The authors treat as sensitive private enterprise data. The model has two layers: the first layer discriminates between non-enterprise data and the other documents (deleting the first type of data), while the second one distinguishes between public and private documents. They use a Support Vector Machines (SVMs) model [30] based on unigrams with binary weights, so they assume that the word sequences differ from one document to one another. The model has been evaluated on different corpora, created from WikiLeaks, Google private dataset, and Wikipedia dataset (see section 2.5). The classifier outperforms the Naive Bayesian (NB) baseline: it can identify more than 97% of information leaks; it achieves a false negative rate (FNR) of less than 3.0% and a false discovery rate (FDR) of less than 1.0%. In particular, it presents a high false positive rate (FPR) on non-enterprise documents due to common features of overfitting and overweighting.

Islam et al. [78] conducted a study to understand people's behavior relating to privacy on social network platforms, such as Twitter. They collected 426,464 tweets and they annotated 270 users as non-private (users 1), slightly private (users 2), or private (users 3), in relation to their shared context (see section 2.5). For the classification task, they used AdaBoost [46] with NB classifier [96], while Latent Dirichlet Allocation (LDA) [18] has been used to discover topics. The results for the binary classification task (users 1 and 3) achieved the 95.45% of acc., while for the 3-class classification task (users 1, 2, 3) the model obtained 69.63% accuracy.

Word embeddings. The theory of word embeddings or distributional theory, [66] considers vector space word representations, where the word vectors are closer if the words occur in the same linguistic contexts i.e. if they are semantically similar. They are traditionally extracted through n-grams or supervised learning. Frequently we talk about static word vectors e.g., Word2Vec [110], Glove [133], and dynamic word vectors e.g., ELMo [134] and BERT [34] (BERT is deepened in the next paragraphs 4.2). Due to the low dimensionality that characterizes word embeddings, they are able to detect the semantic qualities of terms and their relations in a more specific way and to capture the semantic relations of terms within a collection. The dynamic word vectors, compared to the static ones, can distinguish different word vectors in relation to the context where the word appears. Dynamic word vectors are very interesting for sensitive information identification which aims to be context-based.

However, we list here the works based on static word vectors in SID, which allows for considering a sensitive fixed context.

Macdonald et al. [104] present work to protect sensitive government documents using word embedding features for sensitive classification. They state that since the identification of sensitivity is not a topic-oriented task and indeed tends to arise as a product of specific factors, classification can be a good way to face the issue. They also argue that n-gram approaches are too limited for the complexity that characterizes texts with sensitive content: for sensitivity classification on complex document structures, they expect larger values of n to be more effective. They tested their model using a dataset of 3,801 government documents labeled as sensitive or not by experts. They define a baseline where n-grams, part-of-speech (POS) or semantic features are not considered. The best performance is achieved when they extract word embeddings n-grams, word embeddings POS and semantic features combined. Their model achieves a 0.54 F2 and outperforms the baseline out of 9.99%. At the same time, they compare semantic features with grammatical features derived from POS tags and n-gram features.

Word embeddings have been explored also in a paper dedicated to the measure of the sensitivity degree [12]. The authors point out three critical aspects in privacy detection: the lack of a cognitive and perceptive definition of sensitivity; the lack of sensitive annotated corpora at multiple levels e.g., sentence, topic, and term; the context-aware sensitivity, e.g., the distinction between a sentence about a medical doctor that talks about cancer by telling a joke and a doctor who talks about it with one of his patients. They propose a model called content sensitivity analysis (CSA) which assigns a sensitivity score not only to text but also to multimedia content, such as video or picture, combining sentiment analysis features. They have distinguished a basic CSA for binary classification and a continuous CSA for the continuous multi-class classification one. Furthermore, they reflect on more than one sensitive level: a content-level to define sensitive and non-sensitive sentences, a topic-level, and a fine-grained level to define the type of sensitive content more accurately. They have applied a Bag-of-Words (BoW) and word embedding approach to an annotated corpus of social media posts to classify sensitive and non-sensitive information highlighting the limit they observed. The word embeddings approach is based on GloVe [133]. The dataset is composed of 829 posts from social media, labeled by 14 annotators recruited via crowd-sourcing. The posts are tagged as non-sensitive, sensitive, or undecidable (the latter have been deleted). They run the experiments using different ML classifiers. The word-vector Random Forest (RF) has presented the best performances, even if not satisfactory: 0.67% F1 for the sensitive class and 0.69% F1 for the non-sensitive class. These low results can be due to some corpus problems, such as the high number of false negatives and the low number of tagged sensitive posts (only 221). Furthermore, they believe that BoW or word embedding approach and basic classifiers, such as those used, cannot detect the complexity of the task and suggest considering more complex linguistic and semantic aspects or more sophisticated ML models for sensitive identification.

2.3.1.2 Inference rule-based and ontological approaches

Inference rules. An inference rule is a logical scheme that allows an inference [32]. It can be formalized as a mathematical function, which starts from a set of premises or assumptions to reach a conclusion (‘If p then c’):

$$A \rightarrow B$$

The rules operate on the syntactic structure of the utterances; it does not imply any semantic property. In our case, upon the occurrence of a certain syntactic condition (premise), the model will make a prediction (conclusion). An association rule mining [3] is a rule-based machine-learning method that allows one to make inferences among a large quantity of data. Let $I = \{i_1, i_2, ..., i_n\}$ be a set of binary attributes, called items. Let $D = \{t_1, t_2, ..., t_m\}$ be a set of data, called database. Each t in D contains a subset of the items in I. An association rule can be defined as $X \rightarrow Y$, where $X, Y \subseteq I$.

Association rules use the criteria of support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the data, while confidence defines the number of times the if-then conditions are true. In sensitive association rules, if X is a set of words, Y can be a sensitive nature target (health condition or any other personal information). Despite the general rule, in tasks concerning privacy, the model does not need large support, because it aims to find all the sensitive nature content and fight privacy leaks.

Some works in the SID task have implemented this model.

Chow et al. [26] propose an association rule mining model to detect inferences in enterprise documents to reduce privacy leaks and ensure privacy. The specific investigated domains are healthcare privacy legislation compliance and the protection of a corporation's sensitive information (intellectual property, client data, etc.). They based their work on inferences of co-occurrences of words. They work on a first task, which detects a particular topic by identifying all its inference keywords in a Web document; a second task concerns the classification of a sensitive topic when certain words co-occur with seed words. For the first task the sensitive topics are HIV/AIDS, genetic information, mental health, and communicable diseases; for the second task the sensitive topic 'University of Wharton' has been explored in the Enron dataset (also used for the evaluation test). The model supports not only 'conjunctive' inferences (e.g., co-occurrences of items) but also 'disjunctive' inferences that establish a relationship between a conjunctive set of precedents and a disjunctive set of consequents. The inference model achieves a recall of 81% and a precision of 73%.

Geng et al. [58] propose a framework to measure the privacy content in free text documents; it means not only detecting sensitive documents to uniquely identify a person, but also defining a sensitivity score of the personal sensitive entities. The challenge here is the analysis of free documents, instead of tabular databases. They focus on PHI entities in medical records. They distinguish two types of sensitive information:

- *Quasi-identifying attributes (QIA) or quasi-identifying entities (QIEs)*, information that can be used to identify a person e.g., name, address, gender, age, weight, height, etc.
- *Sensitive attributes (SA) or sensitive entities (SEs)*, sensitive information for a person e.g., diseases, credit rating, religion, etc. The SEs can be both (a) objective, or (b) subjective. We can assume for example the sentence '*Marco has been recognized as having a disability*' as (a), while '*Marco often suffers from migraines*' as (b). While the former statement can be objectively measured, the latter depends on social and cultural factors.

When dealing with the degree of sensitive information, it is also important to define concepts that refer to anonymization methods, because they depend on the distribution of sensitive attributes in each class and the database:

- *k-anonymity* [164], the k-anonimity version of a dataset implies that the individual's sensitive information within that dataset can't be distinguished from other individuals whose information appears in the same dataset. In other words, by combining sets of data with similar attributes, it can't be possible to identify information about a unique individual.
- *l-diversity* [105], l-diversity is a form of group-based anonymization based on preserving privacy in datasets by reducing the granularity of a data representation. It means that a class of sensitive attributes must be significantly represented to be reduced.
- *t-closeness* [97], similar to l-diversity algorithms, treats the values of an attribute distinctly and considers the distribution of data values for every sensitive attribute. A class has t-closeness if the distance between the distribution of the sensitive attribute in that class and its distribution in the whole dataset is more than a threshold t.

The QIEs are identified through NER. To resolve the objective and subjective SEs ambiguity, the authors state some principles that can measure the degree of sensitivity by excluding the attribution of a binary score (0 = non-sensitive, 1 = sensitive) through inference rules.

They specifically use MeSH[2], a medical ontology that records terms for diseases, medications, procedures, etc., and shows the relationship between them; they identify 5 diseases to measure. Finally, they evaluate the model identifying PHI in medical discussions from the Web. In free text, a practical problem is a correlation between QIEs and SEs. With the inference-rule-based model, these correlations can be perfectly identified.

Even in Cumby et al. [31], the authors propose a so-called Text Inference Control method to detect and sanitize sensitive information using a multi-class classification framework. They state the definition of k-confusability, which means an adversarial classifier that generates an ordering of the possible sensitive classes of a document and defines as k-confusable a document if the rank in the ordering is at least k. They show experimental evaluation in publicly available datasets.

Pointwise mutual information (PMI). PMI is a statistical measure of association. Mutual Information (MI) in linguistics has been introduced in lexicography [27]. It is very used in this field because it can find the probability of collocations, associations, and similarity degrees between words. MI is measured given two words x and y, and a probability *P(x)* and *P(y)*. MI is defined as:

$$MI(x, y) = log_2 \frac{P(x, y)}{P(x)P(y)}$$

The probability to observe x and y together (joint probability) is compared to observing them independently (chance). A high joint probability means a genuine association between the two entities. MI refers to all possible instances of random variables x and y, while PMI refers to a single event and measures the amount of overlap between two entities.

Sánchez et al. [148] propose a privacy model, called C-sanitized, for document redaction/sanitization. The model detects the semantic inference/disclosure of sensitive entities in unstructured documents, measuring the association between sensitive and non-sensitive words in a document. In C-sanitized the IC information corresponds to the data that have to be protected. c is the generic sensitive concept to be protected, while the term t is a specialization or a synonym/lexicalization of c e.g., c=Hepatitis t= Hepatitis C or t=immune system disease. PMI is involved as follows: given a document D, the knowledge domain K, and a set of sensitive entities C, the C-sanitized document *D'* is the version that for all c in C does not contain any term t or group of terms T that reveal information about c:

$$PMI(c, t) = log_2 \frac{P(c, t)}{P(c)P(t)}$$

As a single term can contain sensitive information to be protected, also a group of terms potentially neutral if independent may become sensitive if co-occurring:

$$PMI(c, T) = log \frac{p(c, t_1, ..., t_n)}{p(c)p(t_1, ..., t_n)}$$

They finally introduce an alpha parameter, which allows for measuring the degree of sensitivity. The α parameter is a continuous numerical value > 1 that adds a proportional degree of uncertainty to the inferences disclosure. They tested the model on Wikipedia articles, choosing sensitive keywords to be analyzed. With $\alpha = 1$ they achieve very high results in terms of

[2] U.S. National Library of Medicine, http://www.nlm.nih.gov/mesh/ (last access January 18, 2023)

precision, but very lower in terms of recall, while the recall improves when $\alpha > 1$. Anyway, the sanitization with $\alpha = 1$ takes an average of 97%.

In a follow-up work [147], they propose a g(C)-sanitized model, introducing a new parameter g(C) to generalize the entity c and specify the allowed information/semantics disclosure for each c in C. This helps make the model more flexible. The detection concerns specific terms considered sensitive; for this reason, it can be considered context-less.

Ontological approach. Garcia et al. [54] adopted an ontological approach to sensitive information detection. The ontological model proposed in this paper aims to identify associations between potentially sensitive concepts and their subsequent sensitive concerns. The information is not treated in a context-less way but considering its complexity and compositional relationships. The dataset used is a collection of textual information from NASA's James Webb Space Telescope website's 'About section' and 'News section' (NASA/JWST 2016[3]). The sensitive concepts identified included information describing the system, its components, mission, launch, orbit, capabilities, specifications, etc. Sensitive concepts do not correspond to single terms, but correlations of terms that together can equate to sensitive information. They run the text using the NER and the coreference resolution annotator of Standford's CoreNLP system [107]. The information is then transformed into an ontological knowledge graph; subsequently, it can be analyzed through inference, in form of SPARQL queries, to detect sensitivity concerns, present at a document or paragraph level.

2.3.1.3 DL approaches

Convolutional neural networks (CNNs). CNNs [95] are neural networks in which the connections between the artificial neurons are inspired by the organization of the animal visual cortex: individual neurons respond to the overlapping regions that constitute the visual field. The connections are made of multilayer perceptrons; which means that each neuron in one layer is connected to all neurons in the next layer.

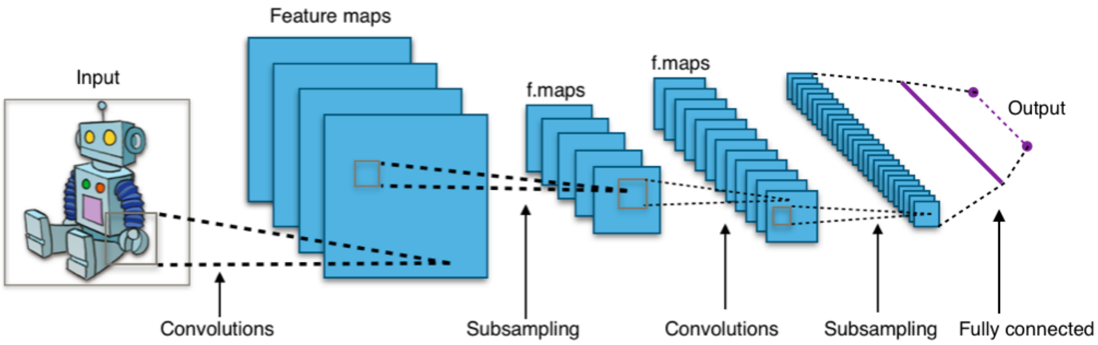


Figure 2.4: CNNs architecture from Wikipedia

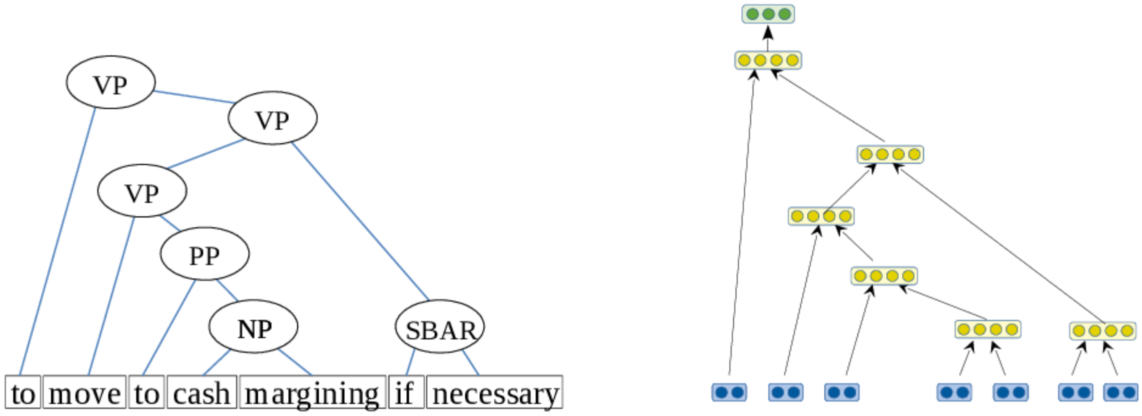
CNNs are mostly used in the image recognition field. Nevertheless, in 2014 a new text-CNNs model has been proposed [83]. The text is processed hierarchically and serialized through multiple layers of word vectors: the source text is continuously reviewed to determine the next output sequence.

CNNs have been implemented mainly in SID works in the Chinese language.

[3] https://www.nasa.gov/open/data.html(last access January 18, 2023)

In Xu et al. [182], military and political domains have been explored. The authors use a Text-CNNs model. The dataset used is made of sensitive Chinese texts from Wikileaks and non-sensitive texts from the Sohu News dataset by Sogou-Lab (a total of about 14,000 documents). The text-CNNs have 4 layers. The results of the models (95.17% acc.) have been compared to the keyword matching approach (73% acc.), the ML detection approach (87% acc.) and the recursive neural network (RNN) detection approach (94.24% acc.). CNNs outperform even the RNNs approaches in terms of accuracy and training speed.

Recursive Neural Networks (RecNNs) and paraphrase approach. A recursive neural network is a deep learning neural network where the same set of weights is recursively repeated in the architecture, in order to produce a prediction over variable-size input structures, by traversing a given structure in topological order. The structure of a RecNN can be a directed acyclic graph (DAG). In Fig. 2.5 we can see a simple RecNN structure.



(a) Parse tree

(b) Recursive neural network

Figure 2.5: RNNs on text from Neerbeck [120]

Concerning the textual analysis and documents as a sequential data structure, some recursive structures such as compositional and constituency dependency parse trees can be added as additional structure information.

RNN on text [120] recursively receives a part of the input structure as input in each step. It presents a tree hierarchical structure, with a root-node and the parse-tree node in each step.

The paraphrase approach considers the possibility to compare two texts verifying, through paraphrasing, if they contain the same informational content [155]. In the aforementioned work, the authors propose a recursive neural network for paraphrase detection. RecNNs help overall in the identification of word order information and avoid confusing sentences, not paraphrased (e.g., '*A cat bit a dog*' is not like '*A dog bit a cat*'). This approach can be applied to the detection of a sensitive document if we consider facing the issue in a contextual-aware approach. In SID, the aim is to compare the sentences to protect with the sentences with which the model has been trained through paraphrases.

The paraphrase approach using RNNs has been used in some sensitive detection works [119, 121]. Neerbek et al. [121] compare some information detection models such as n-gram, Long short-term memory (LSTM), RNNs, and recursive neural networks (RecNNs). RecNNs in most cases achieve the best performances compared to the other models. RecNN model, as the other models, has been tested on the labeled Monsanto dataset and obtained an average of

83% acc. on the silver dataset and an average of 81,7% acc. in the golden dataset. RecNNs have been already explored in a previous work of the author and tested on the Enron corpus [119].

Bi-LSTM approach. The LSTM model [71] is a variant of RNNs. LSTM architecture aims to fill the gap that RNNs present due to their inability to deal with long time lags. This type of RNN can process not only single data e.g., images but also long input sentences e.g., speech. The layers of the network comprise hidden cells and parts of state memory (connected memory blocks).

The basic idea of bidirectional recurrent neural networks (Bi-LSTM or BRNNs) [60] is to consider each training sequence forwards and backward as two separate recurrent nets, both of which are connected to the same output layer. The sequence is analyzed in its standard order and as a reverse copy. The Bi-LSTM model ensures a contextual analysis and more accurate performances. BRNNs have improved the research results overall in speech processing [152].

Due to the accuracy with which they are able to identify the context, the Bi-LSTM models find fertile application in the identification of sensitive data and in fact recently have been adopted by several works in the field.

While Xu et al. [182] have implemented the CNNs for sensitive data identification, we can find in the literature another work in the Chinese language [100] where the authors try to combine CNNs with Bi-LSTM. Their model is a fusion of the two algorithms and is able to obtain localized text features and global text features both. The model proceeds along 3 steps: vector representation, feature extraction, and classification.The domain on which they focused regards unstructured documents containing sensitive information. The neural architecture tries to detect the unstructured documents that contain sensitive information. The datasets used for the evaluation are the same mentioned by Xu et al. [182]. The model outperforms the other models and has obtained a 93.44% of accuracy.

In Genetu et al. [57], the aim is to investigate personal sensitive information in the Amharic language. The authors propose a model to detect and classify them using three types of algorithms: LSTM, Bi-LSTM, and CNNs. They evaluate 7,310 sentences for detection and 6,697 sentences for classification. They create an Amharic annotated dataset. The three sensitive data categories explored are health, political and ethnic origin. The accuracy performances are the following: 82% for LSTM, 87% for CNNs, and 90% for Bi-LSTM, concerning detection; 88% for LSTM, 87% for CNNs, and 93% for Bi-LSTM. So Bi-LSTM is the model which undoubtedly performs better on all tasks.

In Obeid et al. [124], a word embedding-based CNN has been used to explore the clinical text de-identification. They tested both traditional BoW-based ML models and word-embedding-based DL models. The dataset has been constructed by collecting notes from the MUSC Research Data Warehouse, considering a span period of 6 years. The notes have been divided into their section headers using REs, and a group of clinical members labeled them. The labels concern any symptoms of pulmonary embolism. They tested 1,113 illness notes replacing a total of 1,795 HI with the de-identification process. For the de-identification process, they tested an already existing application that uses REs matching and ML algorithms such as Conditional Random Field (CRF) and SVMs. They used BoW-based ML models and word embedding CNNs models for the experimental process. The CNN DL models obtained the best result with an accuracy of 95%. The de-identified dataset lightly improved the performances.

In Guo et al. [64], the Bi-LSTM model is combined with a BERT model. The study is presented in the next paragraph.

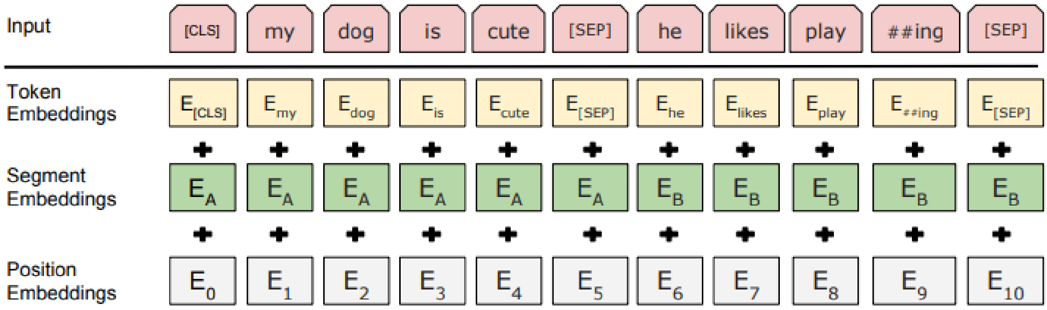

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 2.6: BERT input representation from Devlin et al. [34]

BERT-like approach. The most recent studies on SID have implemented transformer architectures [55, 167, 64]. In particular, the most investigated approach concerns using BERT-like models. BERT (Bidirectional Encoder Representations from Transformers) [34] is a model proposed by Google in 2018 which uses two bidirectional training strategies: the Masked Language Model (MLM) deals with the relationship between words, and the Next Predictive Sentence (NPS) to predict the relationship between sentences. The architecture of BERT is composed of a tokenizer (WordPiece) and a large stack of a transformer, which is provided with the input for training (see Fig. 2.6). The BERT-Base model consists of 12-layer transformers, while the BERT-Large has a 24-layer structure.

So far, this approach has brought very satisfactory results, often outperforming the state-of-the-art. The huge advantages of the pre-trained model are that they do not require particularly large training sets and that they do not require feature engineering for the specific task. Since transformer neural networks have also been used in our work, the structure of the network will be deepened in the dedicated paragraph 4.2; we now illustrate related works.

The pre-trained BERT approach has been used in a study conducted in Spanish [55]. The authors used a BERT-based sequence labeling model to detect and anonymize sensitive data in the clinical domain. Specifically, they use two datasets of medical reports (NUBES-PHI[4] and MEDDOCAN[5]) and they run comparison experiments using CRF and BERT models. CRF is a statistical model composed of a classifier able to make predictions considering the context [92]. It is very used in NLP, such as sequences labeling, POS tagging, or NER. The task consisted of the detection of the sensitive spans and the classification of the sensitive category present in the text span. In NUBES, the pre-trained BERT model outperforms the other systems, while in MEDDOCAN it falls 0.3 F1-score points behind the shared task-winning system, but the authors didn't try more sophisticated fine-tuning layers.

A recent study in the English language [64] proposed ExSense, a model named BERT-BiLSTM-Attention for extracting sensitive information with NLP from unstructured text. The

[4] NUBES is a corpus of around 7,000 real medical reports written in Spanish. Sensitive information has been manually annotated and replaced for the corpus before being published https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus (last access January 18, 2023). NUBES-PHI (NUBES with Personal Health Information) is a version that consists of 32,055 sentences with 11 different sensitive labels. Overall, it contains 7,818 annotations and it is not publicly available.

[5] MEDOCCAN(Medical Document Anonymization shared task dataset) is a synthetic corpus of clinical cases enriched with sensitive information by health documentalists https://temu.bsc.es/meddocan/ (last access January 18, 2023). It contains about 1,000 documents and 23,000 annotations with a total of 21 sensitive categories.

experimental process is conducted on the Pastebin[6] dataset, manually labeled with personal information. Personal information refers to identifiable persons, such as name, address, date of birth, Social Security Number (SSN), and telephone number (see section 2.5). ExSense presents two main modules:

- A content-based sensitive information module to extract REs with predictable patterns;
- A context-based sensitive information module based on automatic ML extraction. The sensitive information extraction is none other than a sequence labeling problem and it deals with the BERT-BiLSTM-Attention model.

The attention model is based on layers able to detect the most important semantic information in a sequence, avoiding focusing on all the information [186]. The performances of the model have been compared to some baseline models. This model performs with an F1 score of 99.15%. As the authors state, a limitation of this study is that ExSense can identify limited types of sensitive information.

Timmer et al. [167] experimented with the BERT model on the Monsanto dataset. They use the BertForSequenceClassification bert-base-uncased pre-trained model (a MLM model). This model does not distinguish between lower case and upper case letters which is a common approach for English NLP tasks; they use a 12 layers model. They achieve an average of 84% of accuracy on the silver Monsanto dataset. In terms of F2 score, they outperform the existing sensitive identification models. The results highlight also a strong dependency on random seeds chosen during the fine-tuning phase. Anyway, the authors conclude by recommending the use of fine-tuned pre-trained transformer models in particular for sensitive information detection in the industry.

2.3.2 Other hybrid approaches

Some recent studies, although not as numerous, have tried to combine more traditional ML or inference-rule methods with DL neural network approaches. The hybrid approach will also be the direction of our work.

A study conducted on detecting basic personal information in the Portuguese language that combines several techniques can be cited [35]; it involved rule-based methods, ML, and neural network models. In particular, the authors explored the CRF algorithm and a Bidirectional-LSTM approach, already cited in Timmer et al. [167]. The task concerns NER and its hybrid architecture can be seen in Fig. 2.7.

Potentially sensitive documents are used as raw text: the model aims to recognize and classify sensitive data, among its classes e.g., personal identification number, socioeconomic information, etc. Firstly, the rules-based model tries to identify NER, considering the context of a specific word or a set of words for each class of entity that must exist in the text. Secondly, a lexicon-based model combines the morphological analysis results with the techniques of stemming and lemmatization to recognize the sensitive entity classes, based on the DataSense NER Corpus. Finally, the last layer is composed of the ML and DL methods. On one side, the statistical models mostly used in NER, such as CRF and Random Forest, are compared; on the other side, Bi-LSTM neural network model has been tested. The corpora used to test the models are the

[6] https://pastebin.com (last access January 18, 2023)

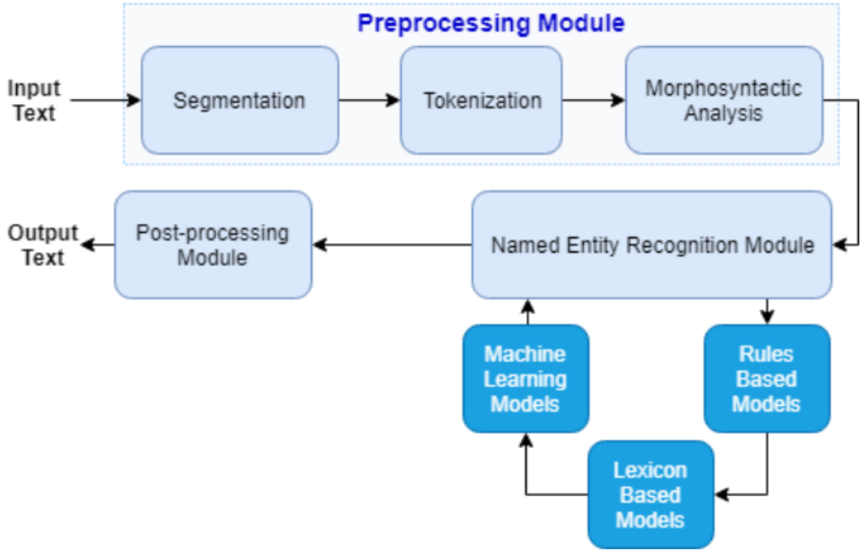


Figure 2.7: Hybrid approach for Portuguese NER in Dias et al. [35]

HAREM golden Collection[7] and the SIGARRA News Corpus[8]. For what concerns the statistical models, CRF is the model that obtained the best results (65.50% F1), while the Bi-LSTM model achieves the 83.01% F1.

To sum up:

- The related works on SID task differ considerably from each other, in particular as regards the domain of investigation, language, the context of identification, and the type of approach. A resume of the studies conducted to address the SID task can be found in Appendix A.1;
- The type of approach could be no context-aware or context-aware and the issue can be faced through different techniques: NLP and ML approaches, inference rule-based, and ontological approaches, DL approaches, hybrid approaches;
- In literature, the NLP and ML approaches concern mainly NER, n-grams, and word embeddings;
- The logical and symbolical approaches concern inference rule-based methods, PMI, and ontological approaches;
- The most recent DL approaches use CNNs, RecNN, Bi-LSTM, and BERT-like approaches;
- There are a few examples in the literature of hybrid approaches.

[7] The only freely available Portuguese dataset annotated with classes of entities was the one developed for the HAREM events.

[8] A dataset from the SIGARRA information system at the University of Porto annotated for named entities (905 news manually annotated) [149]

2.4 Privacy Ontologies

What is an ontology? '*An ontology is an explicit specification of a conceptualization*' [62]. The concept of ontology in informatics comes from the philosophical concept of '*ontology*', as the discipline that concerns the organization of reality. Entities can be grouped, organized within a hierarchy, and divided according to similarities and differences. An ontology aims to construct a knowledge-based system in order to explicitly represent what exists. In these terms, an ontology is an abstraction, a meta-level conceptualization of a specific domain, where entities and relationships are represented.

Ontological models based on such philosophical foundations, but not focused on the metaphysical approach, have been adopted in the fields of Computer Science, Artificial Intelligence, and the Semantic Web. Ontologies have been used and designed as a formal substrate of the Semantic Web, through the Web Ontology Language (OWL). The language used for designing ontologies in the Web of Data is the Web Ontology Language (OWL).

In Guarino [63] seven definitions of ontology are discussed. Following the distinction proposed by Visser and Bench Capon [176], we can highlight three ontological aspects or commitments: (a) task commitments are ontologies with a precise task to solve e.g., an ontology for a diagnosis task; (b) method commitments concern the specific method perspective on the domain knowledge; (c) domain commitments are ontologies referred to a particular domain. Furthermore, according to the authors, legal ontologies have been mainly used for knowledge acquisition (describing entities and relations of a specific domain) and system design (ontologies as reusable construction in the design of knowledge systems) [170]. The authors identify also a genericity degree: upper-level ontologies are characterized by high genericity, while less generic ontologies can be called application ontologies. Van Heijst [172] distinguishes ontologies according to two dimensions: (a) the amount and type of structure of the conceptualization; (b) the subject of the conceptualization. For (a) we can highlight terminological ontologies, information ontologies, and knowledge modeling ontologies. For (b) we can see generic ontologies, domain ontologies, application ontologies, and representation ontologies.

Concerning privacy ontologies, as Palmirani states [127], numerous privacy ontologies can be found in the literature (e.g., HL7 for eHealth, PPO for Linked Open Data, OdrL for modeling rights, etc.). Nevertheless, these can be considered domain ontologies, while application ontologies are missing. In the next paragraph, some application privacy ontology recently developed can be discussed. Finally, we present an authoritative resource we adopted as a reference, the Data Privacy Vocabulary: a top-down ontology realized to guarantee interoperability among the privacy concepts.

Privacy ontological and taxonomic resources are very useful in our work because they allow us to start from a hierarchical and validated organization of a concept as broad and controversial as that of privacy.

2.4.1 Overview

SPECIAL (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance)[9] is a European H2020 project that aims to provide technical solutions for data protection requirements associated with use-cases involving big data. The research call in Big data PPP (privacy-preserving Big Data technologies) started in January 2017 and lasted for three years,

[9] https://specialprivacy.eu/ (last access January 18, 2023)

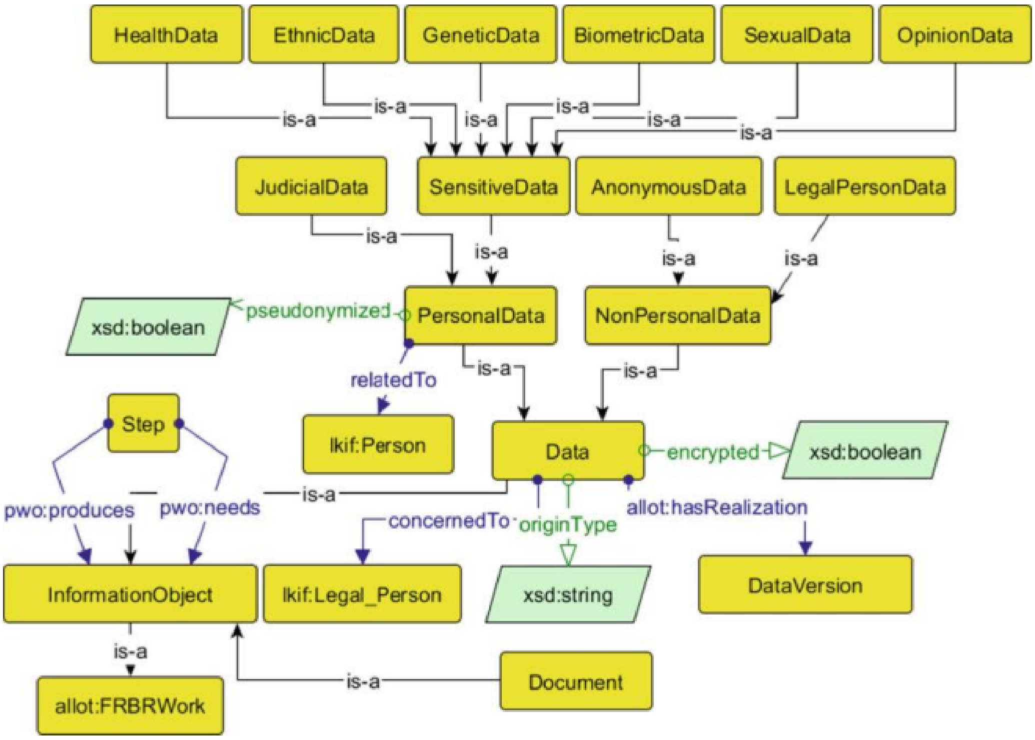


Figure 2.8: PrOnto: document and data model from Palmirani et al. [127]

with the scientific direction of Prof. Andrea Bonatti.

SPECIAL allows citizens and organizations to share more data while maintaining control of the same, thus creating a relationship of trust between the entities involved and at the same time providing information and knowledge useful for the creation of innovative services. SPECIAL uses a ‘Usage Policy’ OWL2 ontology [188] and existing vocabularies, e.g., DCAT[10] and PROV-O (a W3C resource to represent provenance information) [72].

PrivOnto [125] is an annotated semantic framework to represent privacy policies and it has been developed in collaboration with privacy experts. 23,000 annotations have been extracted through SPARQL queries from 115 privacy policies.

PrOnto [127] is a first draft privacy ontology for supporting researchers and regulators while analyzing privacy policies through SPARQL queries. PrOnto is developed following the methodology of MeLOn (Methodology for building Legal Ontology). The aims of PrOnto were (i) to model data protection legal norms starting from legal texts but including also social norms, practitioner opinions, or social behaviors; (ii) to build a legal ontology that is usable for legal reasoning; (iii) to build a legal ontology usable for web of data and information retrieval. PrOnto consists of different modules: documents and data, actors and roles, processing and workflow, legal rules and deontic formula, purposes, and legal bases. Data are defined in GDPR categories, as you can see in Fig. 2.8: PD, non-PD, anonymized data, and pseudoanonymized data.

The GDPRtEXT[11] [128] is an ontological resource that aims to provide linked data to refer to and use text and glossary of the GDPR. It allows the creation of other independent resources that refer to a text or a concept of the GDPR. It has been realized using the European Legislation

[10] https://www.w3.org/TR/vocab-dcat-2/ (last access January 18, 2023)

[11] GDPRtEXT is an open resource: https://openscience.adaptcentre.ie/ontologies/GDPRtEXT/docs/ontology (last access January 18, 2023)

Identified (ELI) ontology to express the legislative points and SKOS to define concepts of the GDPR. In a RDF dataset, every article or point of the GDPR is marked with a URI . This resource fills the gap in methods to address specific sections of the GDPR, consistently linking them.

GDPRov[12] is an ontological resource that, as the name suggests, describes the provenance of data and consents life cycles using concepts and terminology of the GDPR. It is an OWL2 ontology that extends PROV-O and P-Plan [56] (an ontology that extends PROV-O to describe abstract scientific workflows as plans and link them to their past executions) to model the provenance. Provenance is information about entities, activities, and people (or software) involved in producing data. It aims to provide representations of *ex-ante* and *ex-post* activities regarding PD and consent for GDPR compliance; it uses GDPRtEXT to define the concepts of the GDPR. GDPRov contributes to expanding the use of provenance to represent plans or templates to indicate an association between activities in the *ex-ante* and *ex-post* phases of GDPR compliance.

GConsent[13] [130] is an ontological (OWL2) resource that focuses on a particular aspect of the GDPR, the given aspect of consent provided by the data subject. It fills the gap in the model around this type of concept. GConsent aims to model the context, state, and provenance of consent as an entity. It distinguishes between valid and invalid states to use as legal basis for the PD processing and demonstrating the modeling of provenance for activities and agents (such as third parties) and their role in the consent. It uses PROV-O and its GDPR-specific extension, GDPRov (to model the provenance of the consent), and GDPRtEXT (to link concepts to the relevant text within the GDPR).

2.4.2 Data Privacy Vocabulary (DPV)

This section presents a taxonomic reference resource for the representation and processing of personal information: the DPV. After an in-depth description of the resource, we will conclude by highlighting the importance it has assumed for our work. Numerous works in the literature have used or mentioned DPV from its inception to today[14].

2.4.2.1 Introduction

The DPV[15] [129, 131] is a resource created by the World Wide Web Consortium (W3C)[16]. The W3C Data Privacy Vocabularies and Controls Community Group (DPVCG)[17] was formed in 2018 through the SPECIAL H2020 Project (see section 2.4) and aimed at ensuring the interoperability of data privacy through contributions from various stakeholders across computer science, IT, law, sociology, philosophy, industry, policy-makers, and activists. Due to the DPVCG, the DPV can be considered an evolving resource. The community carries on updating and enriching it. The last release dates back to December 2022. It acts as a framework of common concepts and it aimed to fill the lack of the following aspects:

1. Validated vocabularies to represent information about PD use and processing;

[12] GDPRov is available online: https://openscience.adaptcentre.ie/ontologies/GDPRov/docs/ontology (last access January 18, 2023)

[13] GConsent is an online resource: https://w3id.org/GConsent (last access January 18, 2023)

[14] You can find a list of papers that used DPV as a resource, a reference, or that simply mentioned it, here https://www.w3.org/community/dpvcg/wiki/Adoption_of_DPVCG (last access January 18, 2023)

[15] https://w3c.github.io/dpv/dpv/ (last access January 18, 2023)

[16] https://www.w3.org/ (last access January 18, 2023)

[17] https://www.w3.org/community/dpvcg/ (last access January 18, 2023)

2. Taxonomies that describe purposes of processing PD which are not restricted to a particular domain or use case;

3. Machine-readable representations of concepts that can be used for technical interoperability of information.

The DPV starts by representing the GDPR, but it aims to act as a core framework of 'common concepts' that can be extended to represent specific laws, domains, or applications. It is developed in SKOS[18] and OWL[19] serialization. The semantics of the resource object and relationships varies according to the type of serialization. We refer in particular to the OWL serialization, using the terminology you can see in Table 2.1.

Concept	**[DPV-OWL]**
Concept	`owl:Class`
is subtype of	`owl:subClassOf`
is instance of	`rdf:type`
has concept	`owl:ObjectProperty`
Relationship domain	`rdfs:domain`
Relationship range	`rdfs:range`

Table 2.1: DPV-OWL semantics

As the aforementioned resources, DPV aims to represent information for legal compliance, for example with GDPR. However, while the other resources present a particular granularity of representing information and refer solely to GDPR, the DPV is a more high-level ontology, with generic legal terms and does not refer exclusively to the GDPR. In other words, DPV models concepts to define PD handling in the legal domain, specifying the purpose, processing, technical, and organizational measures.

2.4.2.2 Resource Structure

The 'Basic Ontology' describes the first level classes, that define a legal policy for the processing of PD (see Fig. 2.9) and represent information regarding the what, how, where, who, and why of PD and its processing. The top abstract concepts are the core vocabulary within DPV and each core is independent of the other ones.

DPV defines a new core concept: Personal Data Handling. It represents how PD concepts can be related in a particular context. Every concept is provided with relationships and properties.

The DPV proposes a taxonomy of key concepts. The key concepts are the following:

- **Purpose** represents the purpose, the reason, or justification, for which PD is processed e.g., Commercial Research: '*conduct research in a commercial setting or to commercialize e.g., in a company or sponsored by a company*'. Purposes are organized within DPV, based on how they relate to the processing of PD in terms of several factors, such as management functions related to information (e.g., records, account, finance), fulfillment of objectives (e.g., delivery of goods), providing goods and services (e.g., service provision), intended benefits (e.g., optimizations for service provider or consumer), and legal compliance. DPV

[18] https://w3c.github.io/dpv/dpv-skos/ (last access January 18, 2023)

[19] https://w3c.github.io/dpv/dpv-owl/ (last access January 18, 2023)

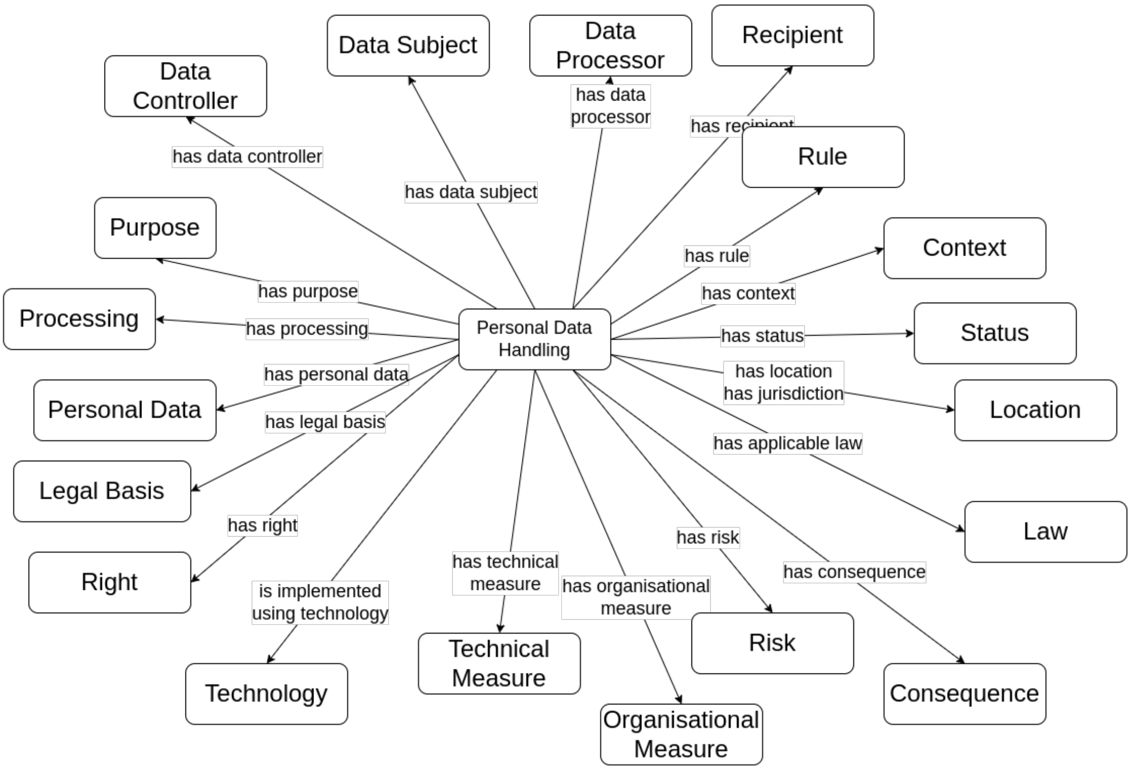


Figure 2.9: PD Handling and core concepts in DPV [194]

provides Sector of Purpose Application that can be used to indicate the relevant information to further clarify or indicate how a purpose should be interpreted.

- **Processing Categories** represent processing in actions or operations over PD e.g., collect, use, share, store. The processing concept is indicated by a source of data (the direct point of data collection or the original/other points where the data originates from) and a processing context (context or conditions within which the processing takes place). Processing Location and Duration indicate where and when it is taking place in terms of location or frequency. Storage and Source of data indicate where data is being stored or will be deleted and where the data is collected or acquired from. For example, data can be obtained from the data subject directly (e.g., given via forms) or indirectly (e.g., observed from activity, or inferred from existing data), or from another entity such as a third party. Finally, Scale, Automation, and New (Untested) Technologies indicate information relevant to the context and the impact of processing; DPV provides concepts for representing whether the processing is carried out on a large scale, consists of systematic monitoring (of data subjects), is performed for evaluation or scoring (of data subjects) - matching and/or combining existing datasets - utilizes automated decision making, or involves innovative use of new technologies.

- **Technical and Organisational measures** represent activities, processes, or procedures used to ensure data protection. Such measures depend on the context of processing involving PD. The concept Technical Measure concerns measures primarily achieved using some technology. Similarly, Organisational Measure represent measures carried out through activities and processes at the management and organizational levels, which may or may not be assisted by technology. An example of a measure can be the pseudo-anonymization and encryption of PD. DPV also provides Policy for representing policies in place and Risk for describing applicable risks and their management or mitigation.

- **Legal Basis** represents a law or a clause in a law that justifies or permits the processing of PD in a specified manner. The legal basis can be customized and applied as needed. DPV provides the following categories of legal basis based on Article 6, GDPR: consent of the data subject, contract, compliance with legal obligation, protecting vital interests of individuals, legitimate interests, public interest, and official authorities. For GDPR, DPVCG provides the GDPR Extension for Data Privacy Vocabulary (DPV-GDPR)[20].

- **Entities** represent the human and non-human actors involved in the processing of PD. DPV distinguishes basically legal entities (human or non-human involved in law), natural persons, and representatives of a legal entity. Entities can be authorities, organizations, data controllers, or data subjects. The Data Controller represents the individual or organization that decides (or controls) the purpose(s) of processing PD. The Data Subject represents the categories or groups, or instances of individual(s) whose PD is being processed.

- **Processing Contextual Information** provides information about the storage condition of data. Location, Duration, Deletion, and Restauration are the concepts that represent this condition, specifying where PD are stored, duration of storage, deletion, or restoration mechanisms (e.g., erasure, or backup availability). Storage information can be part of the processing information (e.g., logs) or technical and organizational measures (e.g., indicating policies or plans in place) depending on context. DPV provides information about automation processing indicating how something is implemented, or who is implementing it, considering also when and how human is involved. Finally, other mechanism of processing data are specified (e.g., use of algorithms, evaluation processing, use of novel technologies).

- **General Contextual Information** provides information about the duration and frequency of an event or operation. It indicates also additional information regarding how the expressed information should be interpreted, or how it applies within a particular context. Context refers to a generic collection of concepts that assists in indicating information such as the necessity, importance, and environment - which aid in the interpretation or application of other core concepts. DPV provides two subtypes of concepts: Contextual Importance and Contextual Necessity (if something is required, optional, or not required). To understand the context it is useful to indicate an identifier for a concept, such as its registration number or internal reference, or other forms of defined names. This permits entities, processing operations, purposes, PD handling instances, and other concepts to be represented along with their identifiers used within a use case or organization.

- **Locations and Jurisdictions** providing relevance to laws, authorities, and location contexts.

- **Risk and impacts** for risk assessment, management, and expression of consequences and impacts associated with processing.

[20] https://dpvcg.github.io/dpv-gdpr/ (last access January 18, 2023)

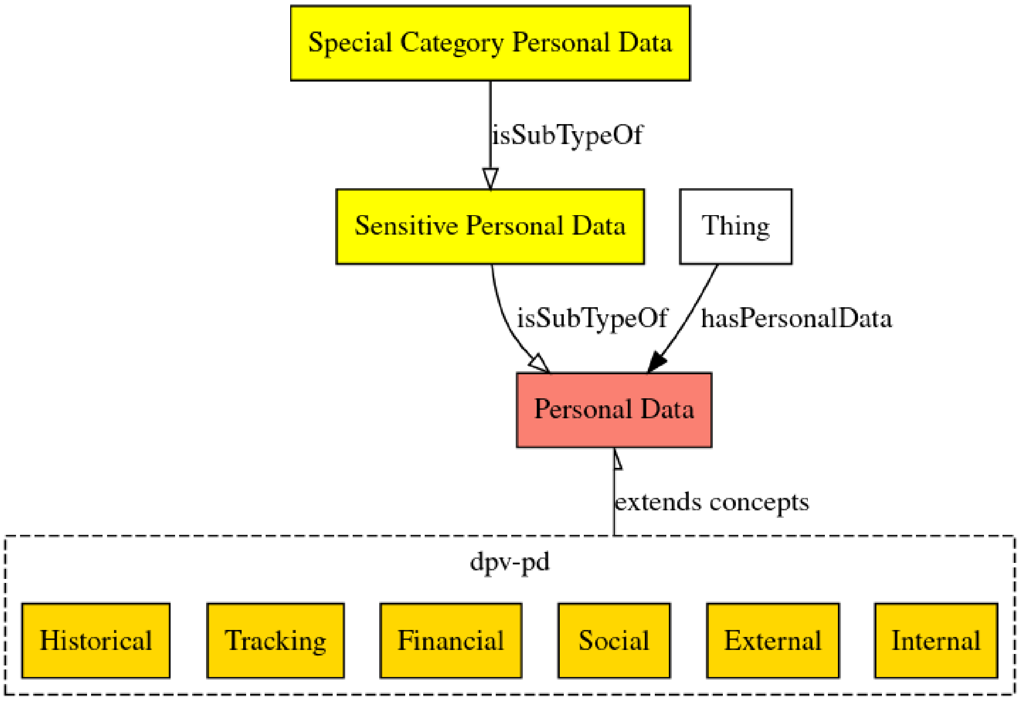


Figure 2.10: PD concepts within DPV and their extension in DPV-PD [194]

- **Rights and Rights Exercise** for specifying what rights are applicable, how they can be exercised, and how to provide information associated with rights.
- **Rules** for expressing constraints, requirements, and other forms of rules that can specify or assist in interpreting what is permitted, prohibited, mandatory, etc.

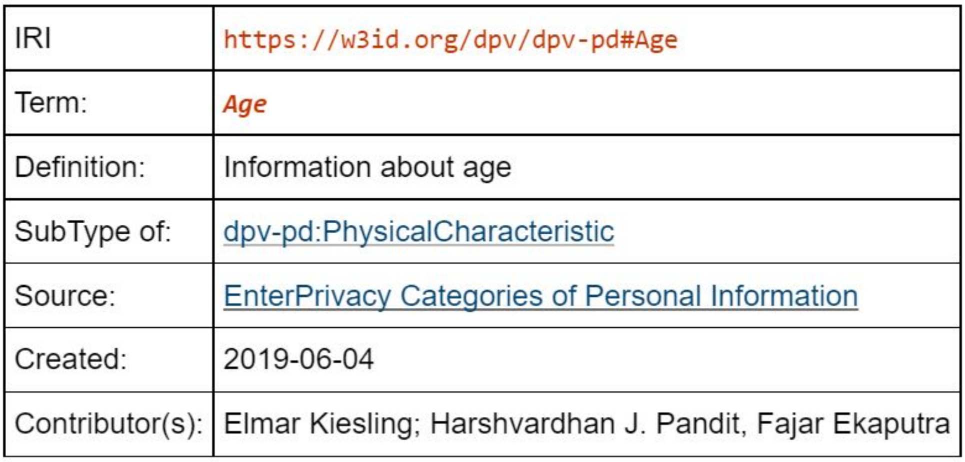
2.4.2.3 Personal Data Categories (PDCs)

As shown, the DPV provides the concept PERSONAL DATA and the relation HAS PERSONAL DATA to indicate which categories or instances of PD are being processed. The DPV has a section, DPV-PD[21], which is an extension that represents a real ontology of PDCs. In DPV-PD the concepts are structured in a top-down schema based on an opinionated structure contributed by R. Jason Cronk from EnterPrivacy (see Fig. 2.10). In particular, SENSITIVE PERSONAL DATA is the class to indicate PD which is considered sensitive in terms of privacy and/or impact, and therefore requires additional considerations and/or protection. The sensitivity of PD can be universal, where that data is always sensitive, or contextual, which means that a use case needs to declare it as such. The SPDs subclass is the SPECIAL DATA category, which includes PDCs such as HEALTH, MENTAL HEALTH, DISABILITY.

Concepts within DPV-PD are broadly structured in a top-down fashion and are divided into macro-categories:

1. **INTERNAL** (within the person): e.g., PREFERENCES, KNOWLEDGE, BELIEFS
2. **EXTERNAL** (visible to others): e.g., BEHAVIORAL, DEMOGRAPHICS, PHYSICAL, SEXUAL, IDENTIFYING

[21] https://w3c.github.io/dpv/dpv-pd/ (last access January 18, 2023)



IRI	https://w3id.org/dpv/dpv-pd#Age
Term:	*Age*
Definition:	Information about age
SubType of:	dpv-pd:PhysicalCharacteristic
Source:	EnterPrivacy Categories of Personal Information
Created:	2019-06-04
Contributor(s):	Elmar Kiesling; Harshvardhan J. Pandit, Fajar Ekaputra

Figure 2.11: Example of representation of a PDC [194]

3. **SOCIAL** e.g., FAMILY, FRIENDS, PROFESSIONAL, PUBLIC LIFE, COMMUNICATION

4. **FINANCIAL** e.g., TRANSACTIONAL, OWNERSHIP, FINANCIAL ACCOUNT

5. **TRACKING** e.g., LOCATION, DEVICE BASED, CONTACT

6. **HISTORICAL** e.g., LIFE HISTORY

There are two additional concepts to this scheme: (i) **DERIVED PERSONAL DATA** and (ii) **INFERRED PERSONAL DATA** to indicate data that are derived or inferred from other PD. In DPV the process of inference or derivation is considered part of the Processing of Personal Data.

The taxonomy currently features 206 categories (classes). Each category is described with: (i) an IRI; (ii) a definition; (iii) hierarchical relations with other taxonomic classes; (iv) the source (the aforementioned EnterPrivacyCategory and DPV Community Group, DPVCG, or SPECIAL project); (v) date of creation; (vi) contributors (see an example Fig. 2.11).

The organization of the PD classes is precisely taxonomic. Analyzing it, we report the levels of the upper classes of the 6 macro categories plus the SPECIAL DATA CATEGORY (Fig. 2.12).

2.4.2.4 Relevance to our work

The DPV has been particularly important for our work, especially for the taxonomic organization it presents and which has been introduced in the previous sections. We have seen how the works in the literature are very different from each other also in the type of domain investigated. Most of the works that focus on the definition of PD refer to PII or to categories belonging to a very specific domain (see section 2.3), such as the medical one. Our work aims to move in the broad domain of PDCs and first requires a valid and structured reference of the area, which only a systematically reviewed resource - such as the DPV - could offer.

The DPV document states that the sensitivity of PD can be universal, where that data is always sensitive, or contextual, which means a use case needs to declare it as such. Likewise, our model aims to cover the identification of universal PDCs. It can be adapted case by case to specific needs, becoming a contextualized model. So we started from the PDCs analysis of the

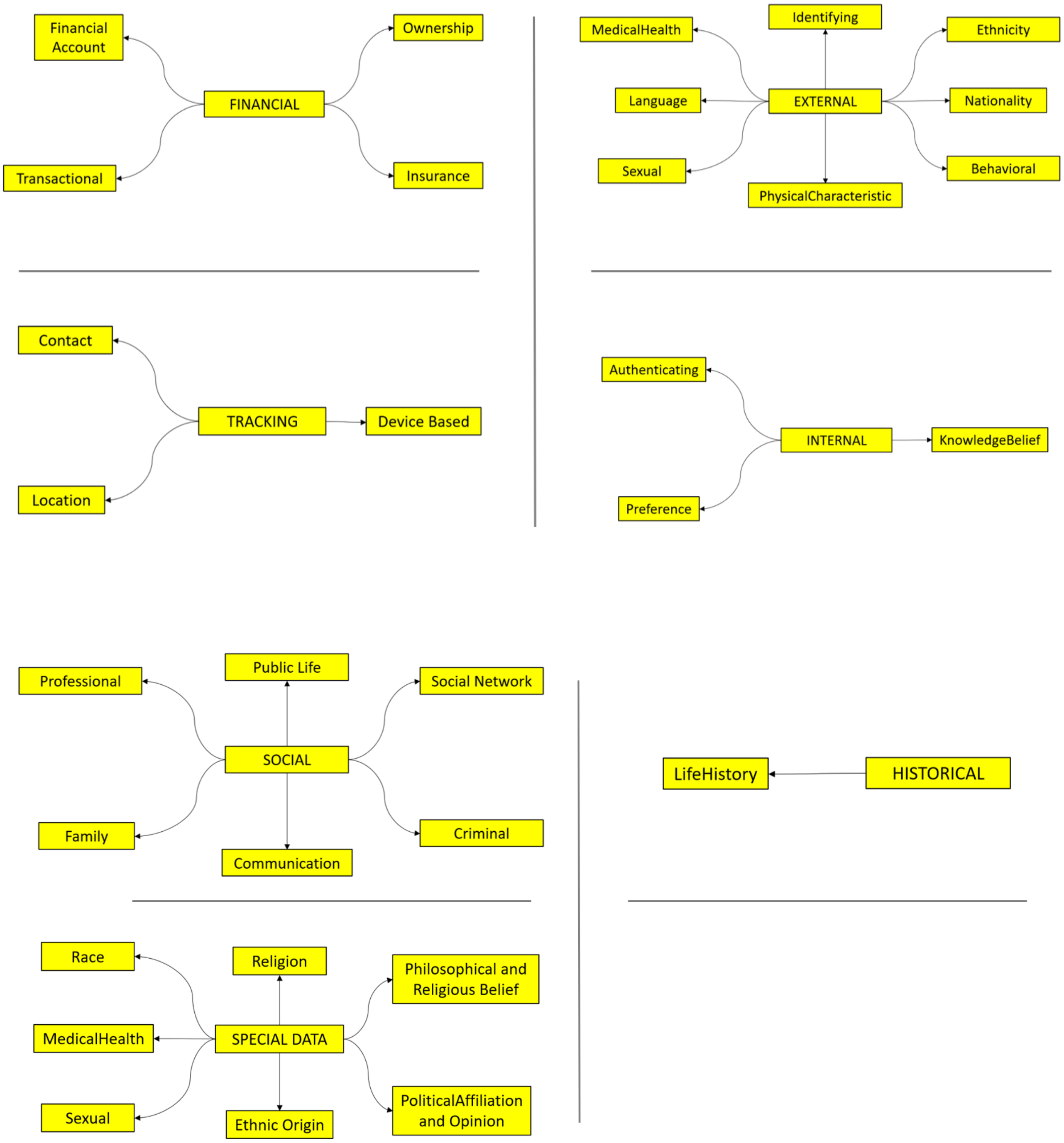


Figure 2.12: Representation of the top-level classes of PDCs macro categories

taxonomy to identify the categories of our interest to investigate. Of course, not all categories have been taken into consideration. For example, categories that referred to non-identifiable data e.g., Browser Fingerprint or Call Log were promptly excluded. The analysis of the feasibility of PDCs will be described in section 4.1.2.

To sum up:

- Ontologies in the legal domain have been mainly developed for knowledge acquisition and system design;
- Privacy ontologies have been mainly developed as domain ontologies. Only the most recent ontologies can be considered application ontologies;
- Some of the most relevant privacy ontologies are SPECIAL, PrivOnto, PrOnto, GDPRtEXT, GDPRov, and GConsent;
- DPV is an ontological and taxonomic resource that describes the Personal Data Handling, pointing out the most important types of PD.
- We have deepened DPV as we have taken it as an authoritative resource for our work.

2.5 Privacy Corpora

A consolidated problem in the field of automatic sensitive data identification concerns the presence of evaluation benchmarks [120]. This is naturally due to the type of processed data, with the consequence that very few examples of labeled datasets with sensitive information are present in the literature. In this chapter, let's go over the few examples presented and used in related work.

2.5.1 Enron dataset

The Enron Email Dataset [85, 86] collects 619,446 emails belonging to 158 users from the American Enron Corporation. It dates back to 2002 and was published in 2003 by Federal Energy Regulatory Commission (FERC) when the corporation failed. It is still publicly available[22] and, to examine real-world fraud, this corpus represents the largest public-domain email database [122]. The Enron corpus was not labeled.

Part of the Enron Corpus (about 2,720 documents) was tagged by human annotators, lawyers, and professionals in a legal track in 2010 [168]. The annotators had to find all the documents about every specific request. They had to take a boolean decision: all the documents have been labeled as true or false regarding the specific request. However, the annotations relate to very specific topics. The specific requests regard for example the following topics:

- The Company's engagement in transactions concerning Real Estate;
- the Company's engagement in prepay transactions (PPAY);
- The Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125);

[22] https://www.cs.cmu.edu/~./enron/ (last access January 18, 2023)

- Financial forecasts, models, projections, or plans at any time after January 1, 1999 of the Company (FCAST);
- Intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard - copy or electronic form (EDENCE);
- Energy schedules and bids;
- Communications between the Company and the financial analysts, conditions, and firms that employ them;
- Discussions about fantasy football, gambling on football, and related activities, including - but not limited to - football teams, football players, football games, football statistics, and football performance.

This corpus was used as an evaluation corpus in some works in literature. In Chow et al. [26], the Enron corpus has been used for the second evaluation task of their rules-based model. They identified a sensitive topic and used the Enron corpus and the Web together to find inferences for the topic. They chose the topic of 'Wharton', the business school of the University of Pennsylvania, and tried to detect the inferences in the Enron documents. The authors state that the private corpus helped them to provide an efficient way to come up with relevant candidate inferences and compensated for the sparseness of the Web.

The Enron corpus has been used also as an evaluation dataset for sensitive document classification [68]. The authors develop several pairs of datasets to evaluate sensitive information detection. The sensitive documents of the Enron dataset (0.6 million documents) have been compared to 581 non-sensitive documents of a corporate website. They included also Wikileaks datasets, and blog posts; for non-sensitive information, they collected, in addition, Wikipedia news articles, Reuters news articles, and text from the Brown dataset. They achieved a high result, however, the difference in the type of source, type of domain, and the unbalanced quantity of documents used for each type of source can be highlighted as the weak point of this work.

Neerbek [120] uses the Enron corpus to evaluate the recNNs paraphrase approach, comparing it with LSTM. The author measured the performance of the Enron and the Monsanto dataset. Regarding the Enron corpus, four topics have been used: PPAY, FCAS, FCAST, and EDENCE. The higher results have been achieved by the PPAY category. The original TREC labels are divided per type of information and have a sentence-level binary label: sensitive or non-sensitive information. A parse-trees labeled sentences resource has been realised[23]. The work underlines how the context and the structure of the sentences help in the SID task.

Kulkarni et al. [91] in order to investigate a personal information domain focused in particular on PII information, using the Enron dataset as an evaluation corpus. They present a clustering-based PII Model (C-PPIM) based on NLP and unsupervised learning to detect PII in the unstructured large text corpus. Even if the Enron dataset contains work emails rather than personal emails, the study aims to protect individual privacy first. Therefore the model was considered only for the content and treated as unstructured text, not for the sensitive information labeled set. The results show as most of the PII is exposed in personal emails: precisely 80%. While personal information is present in 18% of work emails, and only around 2% of forwarded emails.

[23] https://dataverse.harvard.edu/dataverse/enron-w-trees (last access January 18, 2023)

The Enron corpus could be representative of real-world conversation. Nevertheless, since it dates back to 2002, it could not be considered very representative of nowadays communication style. It is particularly relevant with regard to the organizational company domain and confidential information that must be protected. While it also contains personal content, this type of information is not the heart of the dataset.

As we see, the Enron corpus has been used in literature for the evaluation of different tasks:

- The identification of topics through inferences starting from established entities;
- The identification of documents with sensitive content;
- The identification of different types of sensitive data to the dataset labels;
- The identification of PII information through a clustering model.

This is why we can affirm that - up to now - the resource cannot constitute a real shared benchmark in the automatic SID task.

2.5.2 Monsanto dataset

The Monsanto Dataset [187], released in 2017, is composed of 274 legal secret documents. The Monsanto papers have been released to the public as part of the pre-trial discovery process where each party in the trial had access to relevant documents in order to build the evidence for the trial.

Four lawyers annotated and divided it into (i) a silver dataset (labels for weaker sensitive data and noisier sentences) and (ii) a golden dataset, with sensitive data unequivocally labeled. As before, the labels do not affect PDCs; they are of four different types [121]:

- **GHOST**: Ghostwriting, Peer-Review & Retraction. The documents concerning article writing and peer-reviewing by Monsanto paid people;
- **TOXIC**: Surfactants, Carcinogenicity & Testing. The documents concerning discussions and testing of the chemical glyphosate which is part of the Roundup trial;
- **CHEM**: Absorption, Distribution, Metabolism & Excretion. Monsanto articles on Roundup chemistry when Roundup is used in nature. Internal email discussions on Monsanto studies as well as external studies and measurements together with findings, discussions on which questions are answered or need to be answered by Monsanto paid studies;
- **REGUL**: Regulatory & Government. The documents concerning discussions of rewarding people for science that protect Roundup business.

The Monsanto definitely contains sensitive information about the Monsanto company itself and the suing partners.

Neerbek [121] has released a recent version of the Monsanto providing a new golden and silver annotation not more document-level but sentence-level. For the silver dataset, each sentence is labeled with the information type assigned to the whole document by lawyers. If a document contains any kind of sensitive information, all sentences are labeled as sensitive. In other words, if a sentence, that is non-sensitive, appears in sensitive documents, it will be incorrectly labeled as sensitive. The golden dataset is more accurate because it is made of manual annotations from three different annotators. The golden dataset obtained a Fleiss Kappa of score inter-annotator agreement of 0.33; where the Fleiss Kappa metric takes values from negative to 1 and 1 indicates perfect agreement. 118 out of 274 documents have been annotated by humans until now[24].

The authors of the new release of the resource obviously used it also for the evaluation of their recNN model, comparing it with the evaluation of the Enron corpus. In Neerbek et al. [121], the authors have compared four methods (Inference Rule, C-san, LSTM, and recNN) on the Monsanto golden dataset.

The Monsanto golden dataset has been used also as an evaluation corpus by Timmer et al. [167]. The authors have proposed a BERT-based transformer model and have evaluated it against the four Monsanto labels. The results overcome the previous ones on the same corpus (see section 2.3).

This corpus has not yet been used in the literature in large part as the Enron corpus, although it presents strong advantages. It is labeled and has a golden labeled subset; compared to the Enron corpus, the documents in the Monsanto dataset are more recent and the resource has already been reused as a benchmark in the field of identifying sensitive data. However, like Enron, it has a highly specific domain, which is limited to the identification of the four identified labels. If in Enron, a corpus of emails, there are SPD (even if not labeled), Monsanto refers purely to legal documents; therefore the sensitive personal domain can be excluded.

2.5.3 Other corpora in related work

Other studies on SID task have not taken into consideration the Enron or Monsanto dataset; often this happened because of the domain these works intended to investigate and which the aforementioned resources did not cover. We don't deepen here the works conducted on languages other than the currently interested one i.e., the English language. Regardless of the domain, these studies needed to address other resources.

The datasets that have been used in other related works are the following:

1. **WikiLeaks**. WikiLeaks is an international non-profit organization that receives anonymously, thanks to a container protected by a powerful encryption system, documents covered by secrecy (state, military, industrial, banking) and then it uploads them to a personal website. WikiLeaks generally receives government or corporate documents from sources covered by anonymity. Hart et al. [68] use it to create three corpora for their evaluation, identifying the organizations to obtain the documents. They have collected 23 documents from the following sources: DynCorp, a military contractor; TM (Transcendental Meditation), 120 public material from this religious organization; the Mormon corpus that includes a Mormon handbook that is not to be distributed outside of its Members, splitting the handbook into 1,000 character-long pieces and adding two supplemental organizational documents from the church available through WikiLeaks.

2. **Wikipedia articles**. Wikipedia articles or pages are very easy to acquire and contain different types of sensitive information. In Hart et al. [68], the authors have created a Wikipedia test corpus, randomly sampling 10K Wikipedia articles. In Sánchez et al. [148], the aim was to establish a framework for measuring disclosure risk caused by semantically related terms; the authors use Wikipedia pages of individuals (as they did in other previous works [146, 147]) e.g., movie stars. They use a manual annotation for sentences on

[24] https://dataverse.harvard.edu/dataverse/monsanto-w-trees (last access January 18, 2023)

Wikipedia pages relating to sensitive personal information typically defined by keywords and corresponding to PII e.g. HIV (state of health), Catholicism (religion), and Homosexuality (sexual orientation). We believe that creating a corpus based on Wikipedia pages of famous people could be a good method of investigating SPD. Unfortunately, this dataset is not publicly available and, in any case, complex sensitive categories are not considered.

3. **The Google private document dataset**. It consists of 1,119 posts by Google employees to software-development blogs, with which Google collaborates. Hart et al. [68] use also this dataset, treating as private the projects conducted as closed-source development.

4. **A dataset of 1,111 government records** sampled from a larger corpus of documents addressing international relations. The 1,111 records have been manually annotated by government experts with two sensitive labels (from the UK's Freedom of Information Act 2001): (i) the occurrence of personal information (86 out of 1,111) (ii) the material damaging to international relations (27 out of 1,111). This corpus has been used in related works [14, 108]. Due to its sensitive nature, unfortunately, the corpus is not publicly available.

5. **Twitter corpus.** A corpus of tweets and other social information presented in a study by Islam et al. [78], and labeled through Amazon Mechanical Turk (ATM). It counts 426,464 tweets. 500 words from 270 users have been collected and annotated as privacy or no-privacy content. Every user has received a privacy score according to the percentage of private tweets achieved. Even more, the authors trained a sentiment classifier on nine privacy categories: location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family or other associations, personal details, PII, and a not private category that contains objective and neutral tweets. Each category contains at least 6,000 words of training data made up of manually labeled tweets that represent the privacy content. The dataset is interesting even though it dates back to about ten years ago. Furthermore, the study focused on investigating people's behaviors and inclinations to disclose sensitive content on social platforms. Therefore, the annotations are traced back to individual users and their private, non-private, or moderately private behaviors.

6. **Social media posts dataset**. Battaglia et al. [12] propose an annotated social media dataset to identify sensitive and non-sensitive posts. They used a dataset presented by Celli et al. [23], consisting of 9,917 anonymized social media posts. They have chosen 829 posts and they set up crowdsourcing using a Telegram bot: the users had to classify each post as sensitive, non-sensitive, or undecidable. The task has been solved by 14 annotators. The undecidable posts have been deleted and the final dataset is made of 679 posts. However, the dataset has strong limits due in particular to the amount of annotated data that is not sufficient for automatic training experiments and to the non-balanced proportion between the posts labeled as non-sensitive (449) and the sensitive ones (230). The limits of the corpus have been highlighted also in the results of the experiments conducted by the same authors (see section 2.3.1.1).

7. **A PII dataset from Pastebin**. In a recent work [64], the authors have created a dataset collecting data from Pastebin[25], a platform where people can share text and content of different types. Pastebin contains voluminous sensitive data because the users can upload personal information, password credentials, and financial information. The authors

[25] https://pastebin.com (last access January 18, 2023)

collected 1,035,634 documents. They preprocessed the documents, obtaining 144,967 text sequences as training data. They have identified four types of sensitive information in text using REs for content-based sensitive information and the BERT-BiLSTM Attention model to automatically extract context-based sensitive information from the preprocessed text. The sensitive information concerns:

- *Personal information*: name, SSN, date of birth, nationality, address, phone number, occupation, health, education.
- *Network identity information*: IP address, MAC address, Email, social media account, system account, internet, browsing history, chat history.
- *Secret and Credential information*: login password combo, processed password (salt, encryption), encryption negotiation, security question, API key/token, private key, and digital certificate.
- *Financial information*: consuming records, bank account information (account, bank name, bank number), credit card information (card number, CVV, expiration), digital currency (bitcoin, etc).

The dataset used in this work is not currently available. We point out also a limitation of this study regarding the types of sensitive information investigated. The categories refer to PII, frequently detected through REs or very narrow linguistic patterns. This is a limitation highlighted in the conclusions of the same work.

2.5.4 Limits and our contribution

A review of the corpora used in related works on SID task can be found in Appendix B.1. So far, the most used corpus, as it can be seen, has been the Enron corpus. Therefore, there is a clear lack of a common benchmark in the automatic SID task. The main limit - that contributes to this lack - certainly lies in the remarkable difference in the investigated domains among the related works.

Based on the domain of our interest (the PDCs), the labeled corpora are the following:

- Wikipedia corpus, which however presents labeling for keywords referred to a limited number of PDCs. The corpus also collects Wikipedia pages; although it contains personal information, cannot be representative of informal online communication;
- The Government records dataset has a small portion of labeled PD, but the dataset is not public;
- Social media dataset, which has already revealed its limits in the first experimental results due to the insufficient corpus size and the unbalanced quantity of sensitive and non-sensitive labels;
- the Pastebin dataset, which presents labeling on one side for REs of PII and on the other automatically extracted. It is not publicly available.

The Enron corpus, as seen, contains personal information, although its labeling has so far been done on another level. However, due to its release date, it cannot be considered a representative corpus of today's communication.

Our contribution in this sense, as it will be deepened in section 4.1, concerns the construction of a dataset labeled by PDCs that follows the reference taxonomy of the DPV. We have created semi-automatically sentence-level labeled corpora and then the resource has been manually validated by a group of annotators: the first corpus aims to discriminate between sensitive and non-sensitive sentences; the second to identify the different macro-types of PDCs. Finally, the third dataset identifies the fine-grained PDCs. This resource, which certainly could be expanded, can become a benchmark and a training resource for the SID task, with particular reference to the specific domain of PDCs.

To sum up:

- A lack in SID concerns the presence of evaluation benchmarks;
- Corpora used in this task differ a lot, due to the varied characterization of the task described in the previous paragraphs;
- The most used corpora in related works are the Enron dataset, and the Monsanto dataset but neither of them concerns the domain of SPD;
- Other corpora have been used in the literature, but the problem of a point of reference remains, so, in our work, we have focused on this issue.

Chapter 3

Methodology

In this chapter, we establish the theoretical foundations of the methodology of our work. Theories and resources that will be practically used are described here at a high level and in a purely theoretical key. In section 3.1, it will be described a twofold approach followed by the design of the model: a knowledge-driven top-down approach and a data-driven bottom-up approach; we outline the advantages and disadvantages of both approaches.

In section 3.2, another important difference between models is brought to light: semantic-grounded and linguistic-grounded models. When combined, they can create highly promising hybrid approaches that ensure both linguistic and factual understanding of the text. In particular, the use of knowledge graphs for the representation of concepts will be described.

In section 3.3, we introduce the Frame Theory, central to our work, and describe the main resources of our interest. Finally, in section 3.4, we describe the architecture of the transformer models, in particular BERT and RoBERTa. These models were used for the linguistic-grounded and bottom-up approach of the work (Table 3.1).

Methodology	**Contribution**
Frame-based approach 3.3	PRIVAFRAME 4.3
Transformer-model based approach 3.4	Transformer models for SID 4.2

Table 3.1: Methodologies and related contributions

3.1 Twofold approach

In dealing with the work related to the SID task, a further subdivision based on rather general approaches can be formalized: bottom-up approaches on the one hand and top-down approaches on the other. In the bottom-up approach, the features that characterize texts with sensitive content and PDCs are empirically deduced; in the top-down approach, the process is inverse: PD present in the text and the linguistic traits through which they can be extracted are assumed. Another generic difference concerns data-driven and knowledge-driven approaches. Data-driven approaches are mostly those characterized by NLP, ML, and DL methods, using probabilistic and statistical approaches [143]. Knowledge-driven approaches are instead inference rule-based and ontological approaches, using logical reasoning for activity inference [36].

The top-down approach has been extensively explored in different NLP tasks [88, 183, 126]. In SID literature, as seen, top-down approaches are mainly rule-based and addressed through

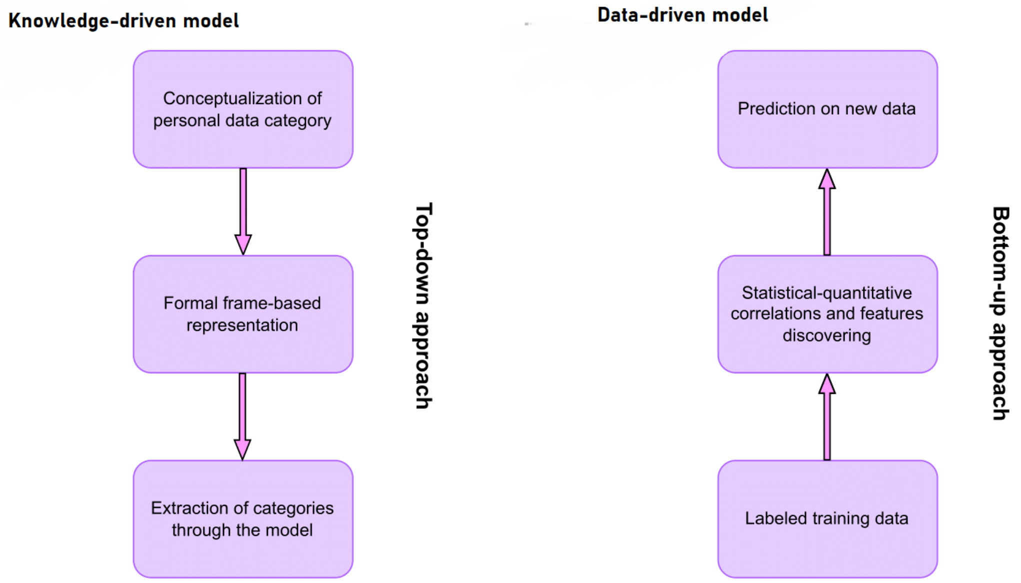

Knowledge-driven model
Conceptualization of personal data category
Formal frame-based representation
Extraction of categories through the model
Top-down approach
Data-driven model
Prediction on new data
Statistical-quantitative correlations and features discovering
Labeled training data
Bottom-up approach

Figure 3.1: Top-down and bottom-up adopted approaches

keywords identification [26] or NER task [58, 54], while bottom-up approaches are based on ML and DL models [120, 55, 64, 167, 182].

What we do is combine a data-driven bottom-up and knowledge-driven top-down approach. The knowledge-driven top-down approach presupposes a high degree of abstraction, as we try to tackle the problem by creating semantic-grounded representations of the concepts of PD, assuming that such representations can help in their pragmatic identification. The abstraction technique is frequently used also in conceptual model design studies [80] and the concept was introduced by Bartlett who observed that people remember their experience through an abstract schema, as a representation of it [10]. Conceptual modeling presupposes the treatment of the complexity of the problem by identifying the objects of the problem and their relationships. This abstraction-based approach can proceed in three different ways: horizontally, vertically, or in a general way. The horizontal abstraction deals with the problem and its various facets, the vertical one deals with the problem broken down into details by analyzing in-depth every detail, while the general abstraction is concerned with putting together the horizontal and vertical partial abstractions to reach a cohesive solution. If on a horizontal level, the complexity of the problem has been faced with the analysis of the various perspectives that emerge from the studies at the state-of-the-art of the previous section 2.3, the abstraction characterizing the top-down approach that we describe here mostly concerns the vertical dimension. We give partial solutions to the problem, starting from the decomposition of the target, the categories of sensitive data, up to its detailed representation; on the other hand, the bottom-up approach of our model is based - as well as models in literature - on a data-driven architecture and the linguistic model stands out as the end product (twofold approach architecture in Fig. 3.1).

A top-down approach based on rules and the conception of a fuzzy processing model based on knowledge constitute the traditional paradigm, often compared with the most recent models based on automatic extraction from data. In Hullermeier et al. [76], the traditional approach is

Knowledge-driven top-down approach	**Data-driven bottom-up approach**
Training data are not required	Data to train the model are required
Deals with high-order problems	The complexity of the problem emerges from the training
Transparency and interpretability	Features discovering models
Granularity	Faster and rawer training
Robustness	Changing data can influence the robustness of the model
Representation of vague patterns	Specific domain-dependent
Highly subjective	Quantitative approach
Manual model adaptation	Less need for manual intervention on model definition

Table 3.2: Characteristics of knowledge-driven top-down and data-driven bottom-up approaches

criticized, when compared to the ML methodology in terms of performance and scalability of the model. However, the author recognizes the interesting opportunity that such models can offer to ML [75].

The advantages that a knowledge-driven top-down approach can offer are primarily the need for training data; if these are essential in a data-driven model, applying a top-down knowledge-based approach they are not and indeed they can be easily and quickly produced. The latter models are also able to analyze, treat and bring out the complexity of the problem that does not emerge in the same way in the statistical-quantitative models; in fact, another characteristic of knowledge-based models concerns the ability to manage granularity. On the other hand, the detailed breakdown of the general problem into the various partial problems can often lead to the generation of a very complex model, while in learning data-driven models we witness the emergence of complexity as a consequence of training. Also for this reason, the former is very suitable for the treatment of high-order problems, such as social or legal problems, and the representation of vague models. A learning data-driven model is not always characterized by a degree of transparency and interpretability, but it is also true that it is characterized by a highly subjective and context-dependent character. A further advantage of knowledge-driven models concerns robustness, as in data-driven models a change in the training data can affect them; while another disadvantage consists in the complexity of manual adaptation. Learning data-driven models, on the other hand, are able to guarantee faster training and feature discovery. The advantages and disadvantages of the two approaches are summarized in Table 3.2.

In particular, the advantages and disadvantages reported in applying the twofold model to our problem are the following:

1. The knowledge-driven top-down method allows facing a high-level problem such as that of PDCs by releasing it from the textual context and providing representation that can be applied horizontally to several specific domains; guarantees a high level of granularity, defining rules for each specific category of data; training is not required. On the other hand, just as it can provide a semantic-grounded representation of the sensitive data categories present in the sentence, it is unable at the same time to provide a strongly context-based interpretation and to work on the binary identification of sensitive or non-sensitive content through text in such a comprehensive way (i.e., considering highly context-dependent tex-

tual borderline cases, such as ironic or hypothetical expressions), as a model trained with a sufficient number of examples could do.

2. The data-driven bottom-up approach guarantees an a priori identification of text with sensitive and non-sensitive content, thanks to highly context-sensitive trained models. On the other hand, more complex multi-classification problems require a large amount of training data, and the model, not starting from a conceptualization but providing experience, is necessarily domain-dependent.

As seen, in addition to the distinction in the type of process adopted, we can theoretically speak of semantic-grounded *vs* linguistic-grounded approaches. We discuss this in the next section 3.2 by introducing a third type of approach in automatic detection and recognition: the hybrid approach. Hybrid activity reasoning combines top-down and bottom-up, data-driven and knowledge-driven methods in order to benefit from the advantages and tackle the disadvantages imposed by each technique. The hybrid model is proposed as the structural framework of our thesis.

To sum up:

- We can identify two types of macro-approaches to solve a generic task: top-down and bottom-up approaches;
- Another substantial difference concerns knowledge-driven and data-driven models;
- All these methods have been tested in the literature concerning the SID task;
- In particular, our contribution presents a knowledge-driven top-down and a data-driven bottom-up approach;
- The knowledge-driven top-down approach implies a high degree of abstraction, which allows us to conceive a conceptual model capable of representing the complexity of PDCs;
- Data-driven bottom-up models are the ones most explored and validated in the literature, but can benefit from knowledge-driven top-down approaches to give the problem a more fine-grained shape and to obtain more accurate identifications;
- The top-down approach we have adopted guarantees a semantic-grounded and fine-grained approach to the problem of sensitive data identification;
- The bottom-up approach adopted allows highly textual context-aware identification and the disambiguation of sensitive sentences from non-sensitive ones.

3.2 A hybrid approach

Recently, transformer models (e.g., BERT, Open AI GPT, discussed in detail in the next section 4.2), have proved to be particularly promising in the automatic processing of language. These models base their assumptions on purely linguistic associations (relationships, patterns, and similarities between the words of the sentences), without taking into account the cognitive or factual basis that such expressions can bring with them. It is in fact for this reason that they are called linguistic models (or LMs) and they differ from cognitive or factual models. As it is easy

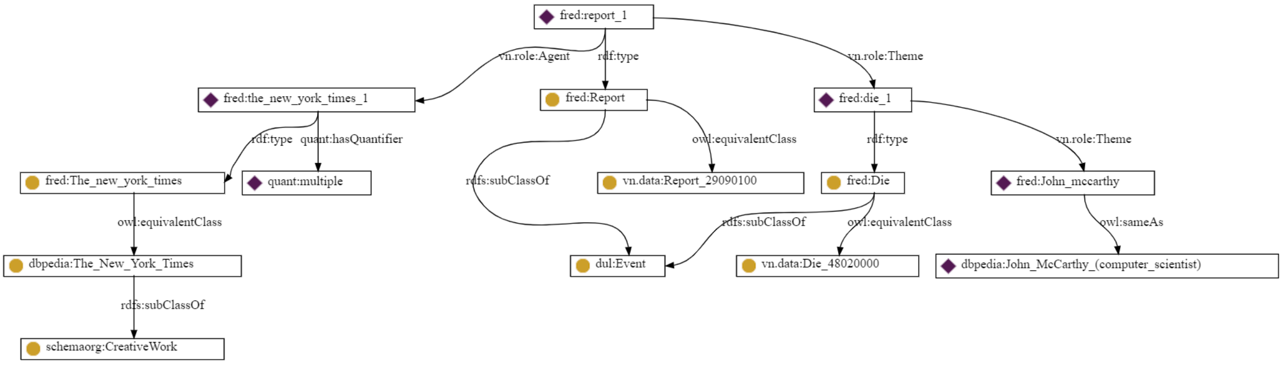


Figure 3.2: Example of an RDF graph extracted by FRED: *"The New York Times reported that John McCarthy died."*

to think, and as introduced in the previous section, LMs are strongly dependent on data and therefore on the specific domain in which they are trained. Moreover, these models are not able to deeply understand the semantics of the sentence, they are therefore not semantic-grounded and consequently not worthy of reliability. On the contrary, cognitive or factual models can provide knowledge unrelated to a specific context and specific training data, due to their natural structure. This feature makes them good candidates for solving higher-order problems such as social or legal problems.

Hence the importance of combining subsymbolic with symbolic models, which can guarantee an understanding that is not only grammatical but also based on knowledge.

A recent project, the Tailor project[1], deals with Trustworthy AI and focuses in particular on a hybrid approach, which combines symbolic approaches related to reasoning and subsymbolic or digital approaches related to learning. The latter still suffer from a lack of explainability, and symbolic systems can compensate for that.

Indeed, it is not enough just to think in extensional terms, at the sentence level, but also in intensional terms to bring out the meaning [52, 53]. The main road in this sense is hybridization: the extraction of linguistic patterns combined with abstraction, through deeper semantic analysis, such as knowledge graphs or ontological relationships.

In the literature, it can be found a practical tool based on the combination of grammatical and semantic theories and approaches: FRED [51]. FRED is a tool for Open Knowledge Extraction (OKE), which produces RDF/OWL ontologies and linked data starting from the Combinatory Categorial Grammar, Discourse Representation Theory, Linguistic Frames, and Ontology Design Patterns. In addition, it can extract NER and Word Sense Disambiguation (see Fig. 3.2)[2]. FRED combines the results of multiple NLP components and formal knowledge representation. FRED was used for the development of Sentilo, a sentiment and opinion extraction tool [140]; Legalo [137], a novel approach to extract links that first identify natural language sentences, and subsequently semantic relations and lexicalizations, and pass it to FRED, which extracts subgraphs; Tipalo [123] analyzes entities based on their definition in natural language and assigns DBpedia links to them. The information is then analyzed by FRED and graphs and word-sense disambiguation are provided. We will use FRED as an extractor for our work.

It is necessary to mention attempts at hybridization which, with an inverse path, try to integrate semantic-grounded theories, such as frame theory, into linguistic models e.g., BERT. The

[1] https://tailor-network.eu/ (last access January 18, 2023)
[2] http://wit.istc.cnr.it/stlab-tools/fred/ (last access January 18, 2023)

reference example is PAFIBERT (Positional Attention-based Frame Identification with BERT) [166][3], where the self-attention mechanisms of the model help to automatically identify frames from FrameNet in natural language sentences. TakeFive [109] is another Semantic Role Labeling (SRL) method, recently proposed, which extracts and links frames and semantic roles. It combines syntactical information extracted with NLP tools such as CoreNLP with Framester, FrameNet, and VerbNet.

The approach we propose to solve the SID task is hybrid and is based on a twofold linguistic-grounded and semantic-grounded model. The symbolic approach in particular is based on the construction of a knowledge graph, allowing a conceptual model that is fine-grained, not domain-dependent, and explainable.

What knowledge graph (KG) means? A knowledge graph could be considered as '*a graph of data intended to accumulate a convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between entities*' [73]. The term '*knowledge graph*' was used for the first time in 1972 [151] in a modular instructional system for courses, but it has become frequently adopted in 2012, when Google introduced its knowledge graph, based on DBPedia Freebase and other several resources[4]. DBPedia[5] and Freebase[6] are open KGs. Another KG based on Wikipedia and WordNet synsets is YAGO[7]. The example of FrameNet can also be cited (deepened in section 3.3).

KGs can be divided following different perspectives:

- Content forms perspective e.g., text KGs, visual KGs, or multi-modality KGs;
- Domain perspective e.g., general KGs and domain KGs;
- Dynamic perspective (according to the timeliness of the contained knowledge) e.g., dynamic KGs and static KGs.

Finally, a knowledge graph construction could imply a top-down or a bottom-up approach. The top-down approach is adopted when the resource is first defined, while the bottom-up approach implies a data-based extraction before a subsequent definition of the resource. Top-down approaches include the following steps: (i) ontology construction, (ii) knowledge extraction, (iii) knowledge fusion, (iv) quality evaluation, (v) knowledge representation, (vi) knowledge storage; while the bottom-up approach involves these steps: (i) knowledge extraction, (ii) knowledge fusion, (iii) knowledge processing, (iv) knowledge representation, (v) knowledge storage [98].

KG extraction means the extraction of concepts and their relations from natural language text, giving them a graph shape of nodes and edges.

A hybrid approach combining a knowledge graph built using OKE with NLP techniques has been adopted also in the legal domain [158]. The study concerns an international private law Regulation and aims to retrieve the relevant legal information through a question-answering system. The authors argue that in legal domain, where we deal with multiple norms and definitions

[3] Unfortunately the paper is still in ArXiv and the code is not yet available.

[4] https://blog.google/products/search/introducing-knowledge-graph-things-not/ (last access January 18, 2023)

[5] DBPedia https://www.dbpedia.org/ (last access January 18, 2023) has been developed in 2007 mainly based on Wikipedia data.

[6] Freebase http://www.freebase.com/ (last access January 18, 2023) is a collaborative knowledge base developed in 2007 and bought by Google in 2012

[7] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/ (last access January 18, 2023)

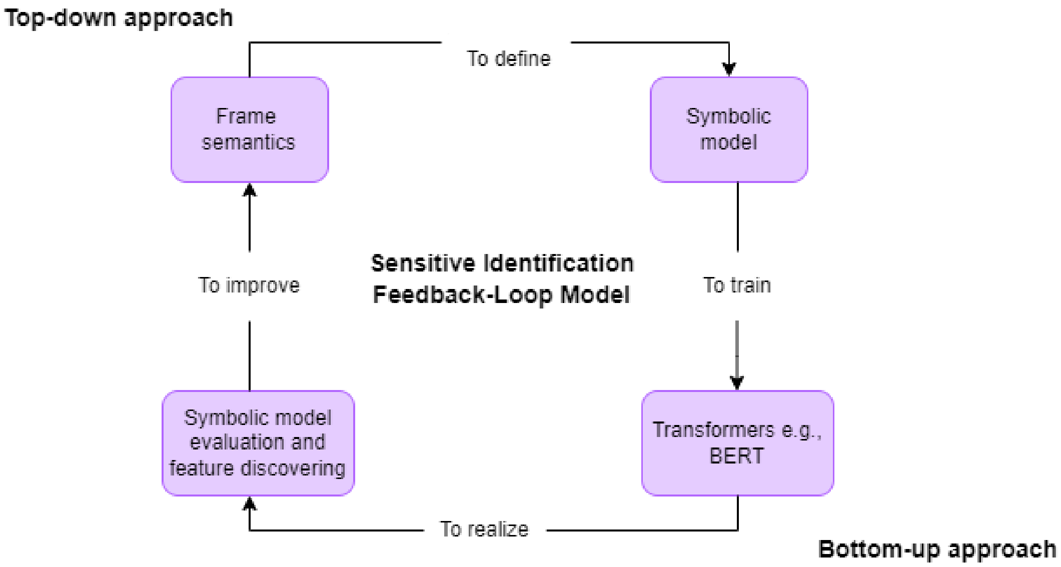


Figure 3.3: Schema of the sensitive identification closed-feedback-loop model

of the same concepts, a KG and a top-down design approach could be a good instrument of representation.

We adopt a top-down design approach for the construction of a KG that could represent the PDCs. The initial design of this hybrid model provided a well-defined bidirectional structure: top-down for the semantic-grounded approach and bottom-up for the linguistic-grounded one, as previously described. However, once the model was implemented, we envisage a feedback loop mechanism that aims at the continuous contamination of the two different approaches (Fig. 3.3). The results of one approach contribute to the redefinition and improvement of the other, in a systematically continuous logic.

To sum up:

- An additional important difference to highlight concerns the semantic-grounded models compared to the linguistic-grounded ones;
- Knowledge-based models can include the cognitive-factual dimension in the interpretation of data;
- Cognitive-factual models can deal with higher-order problems and free themselves from specific training contexts;
- Tools that hybridize linguistic and knowledge-based understanding have been explored in the literature e.g., FRED, PAFIBERT, TakeFive;
- KGs are an excellent and consolidated semantic-grounded modeling approach and their design can be top-down or bottom-up;
- The adoption of a hybrid method can lead not only to a twofold model but to a feedback looping logic in which the results of the one concur to measure the performance of the other.

3.3 Frame-based approach

Our top-down and knowledge-based approach mainly employs Frame Semantics [113] [44]. A frame is '*a remembered framework to be adapted to fit reality by changing details as necessary [...] a data-structure for representing a stereotyped situation*' [113], which draws its origins from the concept of '*schema*' presented by the aforementioned Bartlett [10] and has become of strong interest in the field of AI. Every frame brings with it different semantic information; some of these are always true and necessary for its existence, while others are correlated, defined by the author as '*sub-frames*'. The relations between simple frames could create frame systems: '*these inter-frame structures make possible other ways to represent knowledge about facts, analogies, and other information useful in understanding*' [113]. Fillmore defines the concept of framing as '*the appeal in perceiving, thinking and communicating to structured ways of interpreting experiences*' [43].

The concept of frame has been applied in natural language processing. Fillmore states that the words and the interpretation of an utterance depend on context-dependent experiences [43]. Frame theory is a formal theory of meaning [136] and a solid, cognitively grounded basis that fundamentally contributes to semantic interoperability. A semantic or conceptual frame means the representation of a situation, state, or event through lexical units and semantic roles. Frames are usually evoked by the verbs in the sentence. The meaning of a word can be understood concerning the context by which it is surrounded. In other words, through frames, we can access real-world knowledge. This theory can be applied to the frame detection activity [29], identifying complex relationships in natural language that can contribute to the construction of meaning.

The approach will be based on compositional frames. The abstraction process is frequently based on Frege's principle of compositionality: the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules to combine them. There is no particular evidence of compositional frames in the literature. The term is used in studies conducted on maternal and infant speech, but in somewhat different terms from how we will use it [45]. The authors identify a compositional frame as an AxB structure, in which x corresponds to a precise lexical constraint and A and B to the words that follow and precede it in the expression. What we do is mostly based on the creation of new frames that can represent the categories of sensitive data starting from already existing basic and more abstract semantic frames. The relationship between these frames is created through the use of logical operators. The resource created is a compositional-frame-based resource (see section 4.3). This approach can ensure the following main advantages:

- A semantic-grounded representation of PDCs;
- A context-aware detection of PDCs.

Furthermore, it represents a new approach in the related works regarding the SID task. The association of lemmas to the frames they evoke and to other lemmas belonging to the same frame should help in terms of recognition and affirmation of coherence [44].

For our work, we used Framester with particular reference to FrameNet and WordNet.

3.3.1 FrameNet

FrameNet is a lexicographic project in force at the International Computer Science Institute in Berkeley since 1997[8] [8]: a resource of English based on Fillmore theory [43]. The database contains about 13,000 word senses with examples of meaning; usage and data are freely available.

In FrameNet [145, 8] the meaning of words is described through semantic frameworks composed of frame elements (FEs) that represent an event, a relationship, an entity, or the participants. Each frame presents a name, a description, a list of frame elements with their descriptions and examples (FEs core and FEs Non-Core), and the relations among them.

The main frame-frame relations are:

- **Inheritance**: hierarchical relations where some frames are a more specific version of other parent frames;
- **Perspective**: the same situation can present different frames to specify the perspective they adopt e.g., the frame `Attack` has a relational perspective on `Hostile_encounter`;
- **Using**: a frame that involves other frames e.g., the frame `Defending` implicitly uses the frame `Attack`;
- **Compositional or Subframe**: some frames can constitute subframes of other more generic situational frame e.g., the frame `Invading` is a subframe of `Invasion_scenario`;
- **Precedes**: it expresses the relationship in temporal terms between subframes of the same scenario;
- **Is Causative of**: it provides causative descriptions among frames e.g., the frame `Limiting` is causative of the more generic frame `Limitation`;
- **Is Inchoative of**: it provides inchoative relations among frames.
  E.g., the frame `Change_of_Temperature` is inchoative of `Temperature`.

Other types of relations are expressed as `See_also`.

Finally, the Lexical Units (LUs) are connected to the frame. LUs are words that can evoke this frame. The annotation of the sentences shows how the FEs syntactically adapt to the evoked words. FrameNet has more than 1,000 semantic frames and approximately 11,000 LUs.

For example, the frame `Age` is defined in the following terms: '*an Entity has existed for a length of time, the Age. The Age can be characterized as a value of the age Attribute, or a Degree modifier may express the deviation of the Age from the norm. The Expressor exhibits qualities of the age of the Entity*'. The FEs core are `Age [Age]`, `Attribute [Att]`, `Degree [Deg]`, `Entity [Ent]`, `Expressor [Exp]`. While it presents also Non-Core FEs such as `Circumstances [Cir]`, `Descriptor [Des]`, `Duration [Dur]`, or `Time [Tim]` and LUs such as nouns (age, maturity) or adjectives (ancient, oldish, etc.). The relations concerning the frame `Age` could be seen in Fig. 3.4.

FrameNet annotation data can produce automatic semantic role labeling (ASRL) or semantic parsing. The first automatic annotation model based on FrameNet has been developed in 2002 [59]. Some projects based on ASRL are for example SEMAFOR [90] or Open-Sesame [163].

FrameNet has been developed in other languages, such as German, Spanish, Japanese, and Swedish. A project for the Italian language, IFrameNet, has been also developed [11].

[8] https://framenet.icsi.berkeley.edu/fndrupal/ (last access January 18, 2023)

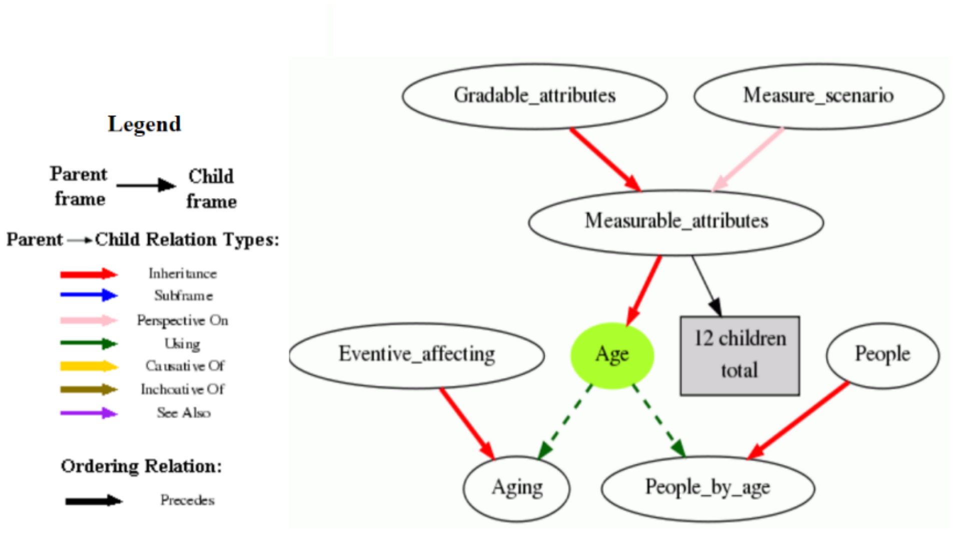


Figure 3.4: FrameNet relations (extraction from FrameNet tool)

3.3.2 WordNet

WordNet[9] [38, 39] is a large English lexical database developed by Miller at the University of Princeton [112]. It contains more than 118,000 synsets i.e., sets of synonyms (nouns, verbs, adjectives, adverbs), each of which expresses a distinct concept, and more than 90,000 different word senses; approximately 17% of the words in WordNet are polysemous containing also compound nouns and collocations. The synsets are interconnected with each other with semantic and lexical relationships. Synsets present semantic relations among them; the main relations are the following:

- **Synonymy** (nouns, verbs, adjectives, and adverbs are involved);
- **Antonymy** (nouns, verbs, and mainly adjectives and adverbs are involved);
- **Hyponymy/Hiperonymy** (nouns are involved);
- **Meronymy/Holonymy**(nouns are involved);
- **Troponymy** (verbs are involved);
- **Entailment** (relations between verbs).

WordNet has been also developed as a multilingual project in more than 200 languages[10], including Italian (ItalWordNet [144]).

Furthermore, WordNet has been involved in several related works. It has been extended and aligned to a formal specification in DOLCE foundational ontology [48] to automatically extract and interpret conceptual relations from it. DOLCE is an upper ontology based on a cognitive and linguistic approach. OntoWordNet is a project which aligns WordNet's upper level with it. The alignment of WordNet and Wikipedia has contributed to the development of BabelNet, a

[9] https://wordnet.princeton.edu/ (last access January 18, 2023)

[10] http://globalwordnet.org/resources/wordnets-in-the-world/ (last access January 18, 2023)

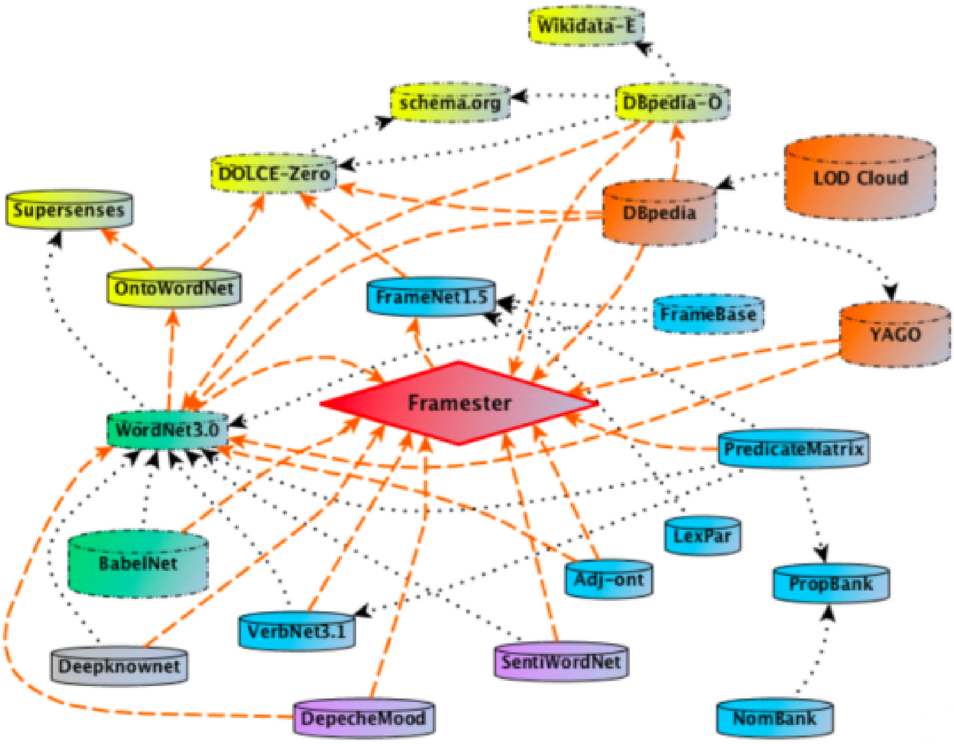


Figure 3.5: Framester structure from Gangemi et al. [50]. Framester is the main hub, while DepecheMood and Sent/WordNet - in purple - are the Sentiment Analysis datasets.

multilingual semantic network [118]. Among others, even FrameNet is linked to the resource, due to their complementary nature [40]. Combining WordNet and FrameNet gives a more complete semantic representation of the meaning of a text than the resources could do on their own.

3.3.3 Framester

The richest knowledge graph containing frame-based linguistic knowledge is Framester [50]. Framester acts as a hub between linguistic resources such as FrameNet, WordNet, VerbNet, BabelNet, DBpedia, Yago, and DOLCE-Zero; it's an interoperable predicate space formalized according to semiotics. Framester uses WordNet and FrameNet internally, expands them to other resources in a transitive way, and represents them in a formal (OWL) version of Fillmore's frame semantics.

Frames are interpreted as multigrade intensional predicates:

$$f = (e, x_1...x_n)$$

Where f is a first-order relation, e indicates the variable for events or states of an affair of the frame, and xi indicates any argument place. Following this definition, in the sentence '*My mum is a medical doctor*', the multigrade intensional predicate is *Be(e, My mum, medical doctor)*; e is the situation, represented in Framester as `FrameClass`. WordNet synsets could be considered as specialized frames or semantic types. They can evoke frames and can be represented in Framester as `SynsetFrame`.

The Framester information about frames is maintained as well as it is presented in FrameNet, but hierarchical relations with a map of generic frame elements and semantic roles are added

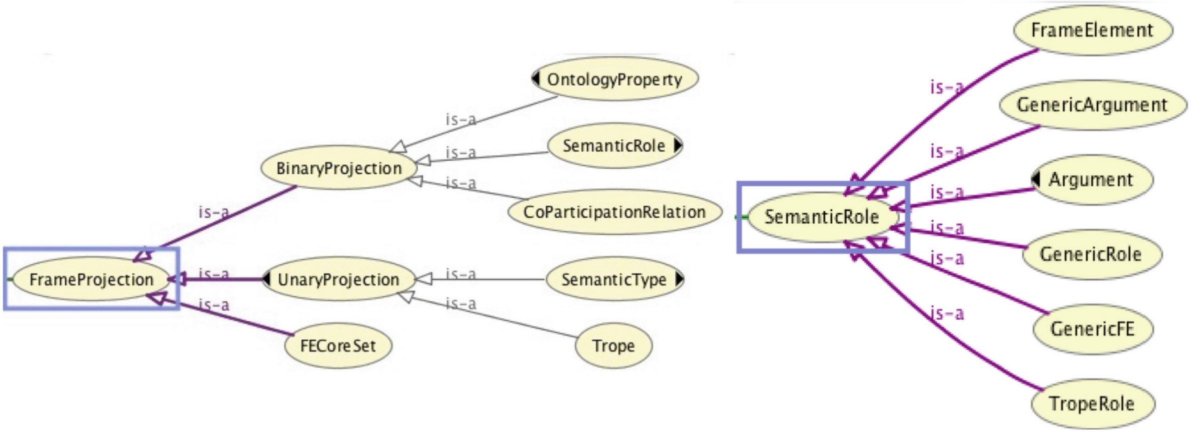


Figure 3.6: Frame projections and semantic roles in Framester [50]

(Fig. 3.6). The semantic relations are created starting from the relations already present in WordNet.

A Framester frame could be retrieved through a SPARQL query from the SPARQL endpoint http://etna.istc.cnr.it/framester2/sparql:

```
 PREFIX fschema:  <https://w3id.org/framester/schema/>
SELECT DISTINCT ?frame
WHERE
?frame a fschema:ConceptualFrame.
LIMIT 10
```

The LUs of the frames link to their `WnSynsetFrame` schema. Due to its expressive potential and the explicit combination of FrameNet and WordNet presented, we have used Framester as a reference resource for our KG.

To sum up:

- We presented the frame semantic theory which constitutes the basis of our top-down approach;
- Specifically, we follow the principle of compositionality applied to frames;
- The main resources we will take into account are Framester, FrameNet, and WordNet.

3.4 Transformer model-based approach

The transformer-based LM is a DL architecture based on self-attention mechanisms introduced in 2017 [175].

The model traces its basis on some previous RNN implementations consisting of an encoder-decoder structure. The RNN Encoder-Decoder structure has been proposed by Cho et al. [25]. The encoder structure receives a sequence $x = (x_1, ...x_n)$ as input, and transforms the information of the first word in a fixed-length hidden state vector of an arbitrary size *h1*. Subsequently, the last hidden state is taken into account to measure the following hidden state of the model. The decoder can create a new sequence as output. It computes a decoder's hidden state input and output, taking into account the previous decoder's hidden state, the previous output, and the context vector. In the end, it produces a $s$oftmax function and a probability distribution.

The Attention mechanism, or global attention, is a technique implemented on a neural network that is based on cognitive attention and has been proposed by Bahdanau et al. [7]; the authors experiment with Bidirectional RNNs, which consists of a forward and backward RNN. The first one works from left to right, and viceversa for the second one; finally, the combination of the two gives the result. It is based on providing a context vector for each decoder step in order to create a model that pays attention to the most important parts of the sequence. The model involves the decoder structure, while a more specific model, the so-called Self-Attention model, tries to consider the encoder stage too.

The Self-Attention mechanism has been introduced in 2017 [99], trying to represent the embeddings and paying particular attention to some parts of the sentence. Giving a sentence S of word embeddings with n tokens $(w_1, w_2, ..., w_n)$, S can be represented as a bi-dimensional matrix, concatenating all word embeddings together:

$$S = (w_1, w_2, ..., w_n)$$

Supposing that each entry of the sentence is independent, we process it in a bi-directional way and we concatenate it in a hidden state. The model should transform a variable-length sentence into a fixed-size embedding:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(w_t, \overrightarrow{h_{t-1}})$$
$$\overleftarrow{h_t} = \overleftarrow{LSTM}(w_t, \overleftarrow{h_{t+1}})$$

The attention mechanism takes the H hidden states as input and outputs a vector of weights a:

$$a = softmax(w_s 2tanh(W_{s1}H^T))$$

The hidden states H are summed up to the weights provided by a (*w1* is a weight matrix with a weight of *da-by-2u*; *w2* is a vector of parameters with size *da*) and give a representation of the input sentence. The vector could represent a specific part of the sentence, as a particular set of related words. To take into account the whole, long, and more complex sentences, they have introduced a multi-head attention mechanism, where the w_{s2} is transformed into a *r-by-da* matrix:

$$A = softmax(w_{s2}tanh(W_{s1}H^T))$$

The resulting vector representation is a sum of A and the hidden states $H(M = AH)$.

However, the RNNs architecture presents some limits concerning the possibilities of parallelization and memory efficiency during training. Due to these constraints, Vaswani et al. [175] proposed a transformer model, where the encoder-decoder structure is composed of stacks of identical and multi-head attention-based layers. The transformer encoder is composed of 6 layers with two sub-layers each: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network; the decoder, composed of 6 layers too, has three sub-layers, adding a multi-head attention layer to the output of the encoder stack (see Fig. 3.7).

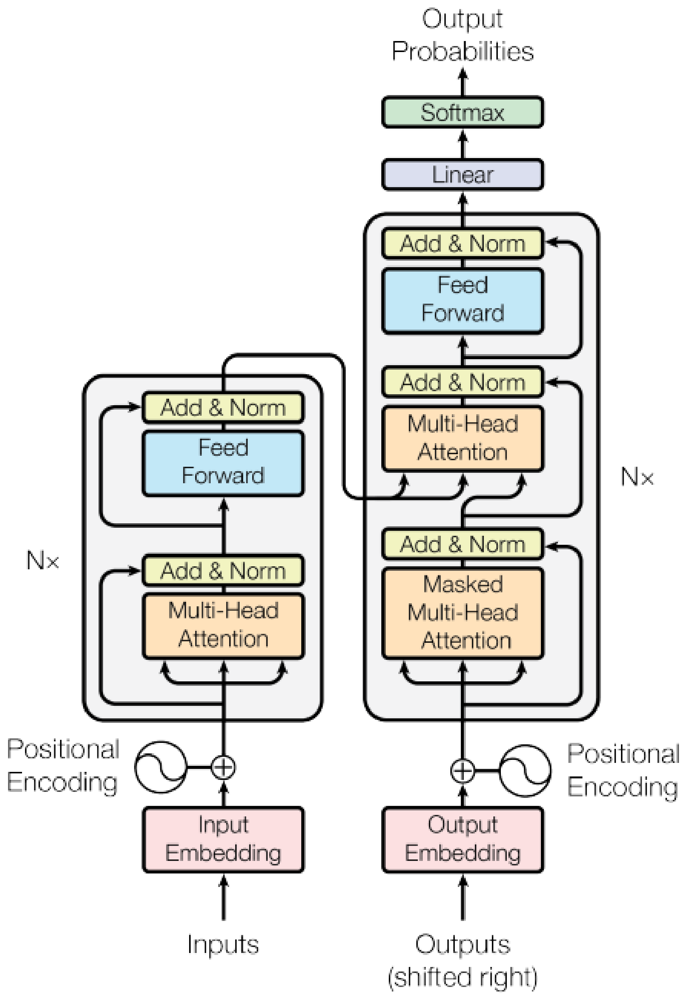


Figure 3.7: Transformer architecture: encoder-decoder structure [175]

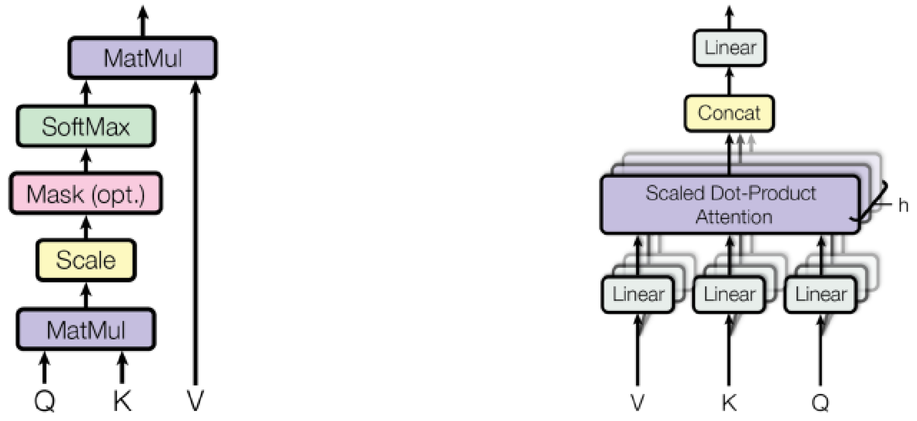
The transformer model is based on two different types of attention mechanisms: the Scaled Dot-Product Attention and the Multi-Head Attention. The first one consists on mapping a query and a set of key-value pairs to output; the second one on learning h linear projection from queries, keys, and values, arbitrarily choosing the h weight and computing the dimension for every attention head in parallel (see Fig. 3.8).

LMs have shown very good results in many NLP tasks, such as paraphrasing, natural language inference, question answering, and NER. In Devlin et al. [34] pre-trained language representations are distinguished into two approaches:

- Feature-based approach, which uses task-specific architectures and where the pre-trained representations are additional features e.g., ELMo [134]. Models like ELMo extract context-sensitive features of each token that correspond to the concatenation of the left-to-right and right-to-left representations;

- Fine-tuning approach, which uses minimal task-specific parameters and is trained fine-tuning all pretrained parameters e.g., Open AI GPT [139].

The main limit of these methods concerns unidirectionality. Devlin et al. [34] introduce the aforementioned BERT to overcome this shortcoming (section 2.3).



(a) Scaled Dot-Product Attention.

(b) Multi-Head Attention.

Figure 3.8: Transformer architecture: scaled-dot-product attention and multi-head attention [175]

The use of MLM allows the model to deal with bidirectionality: the MLM randomly masks some tokens of the input sentence trying to predict the masked words from the context and it is able to combine the left-to-right and right-to-left contexts. Furthermore, BERT presents an NPS that pretrains text-pair representations and lets the model understand relationships between sentences. The BERT model follows two steps: the pre-training and the fine-tuning steps. The model is first trained on unlabeled data on several pre-trained tasks; after that, the parameters are fine-tuned using labeled data from the downstream tasks.

The BERT model architecture has been proposed in two sizes:

- BERT-base model: 12 transformer layers, 12 attention heads, 768 hidden sizes, 110M parameters;
- BERT-large: 24 transformer layers, 16 attention heads, 1024 hidden size, 340M parameters.

A high number of parameters lets the model achieve better performances but increases the memory consumption and prediction time. In 2019, Liu et al. [101] proposed an extension of BERT, which was significantly undertrained: a new model called RoBERTa (Robustly Optimized BERT Approach). The main implementations introduced on the RoBERTa model are the following:

- A longer training of the model with longer sentences: BERT is trained for 1M steps and a batch size of 256 sequences. Training the model with 125 steps of 2K sequences and 31K steps with 8K sequences of batch size improved the task accuracy;
- NSP is removed: the experiment conducted on the new model observes that removing the NSP loss matches or slightly improves downstream task performance;
- The masking pattern applied to the training data is changing: on BERT architecture the train consists of a single static mask. In RoBERTa training data is duplicated and masked 10 times with different mask strategies.

RoBERTa has almost the same architecture as the BERT-model (see Fig. 3.9), but uses a byte version of Byte-Pair Encoding (BPE) as a tokenizer and is pretrained with the MLM task

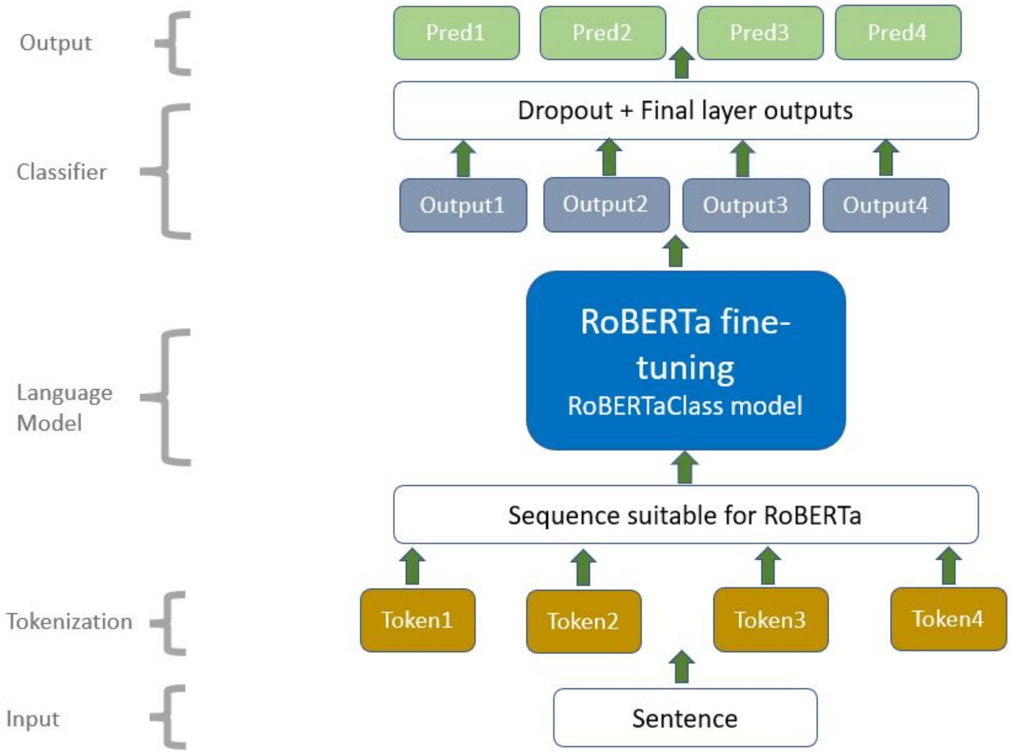


Figure 3.9: RoBERTa architecture

(without the NPS task). It optimizes some hyper-parameters for BERT e.g., longer training time, larger training data, larger batch size, bigger vocabulary size, and dynamic masking. The model outperforms even BERT Large architecture [101] and performs well on different NLP tasks, including classification [21, 138, 16].

Due to the high performances obtained in the literature and the results of the most recent SID task conducted adopting this approach, we use RoBERTa for our classification task. Its high accuracy results led us to consider it a valid DL-based approach for our hybrid model. The experiments conducted and the comparisons with other algorithms are described in chapter 4.

To sum up:

- The mechanisms and architectures of transformer models e.g., the Self-Attention mechanism and the Multi-Head Attention have been described;
- LMs present two main different approaches: a feature-based approach and a fine-tuning approach;
- BERT is a bidirectional encoder representation transformer that recently has established itself as the state-of-the-art, thanks to its MLM and NPS mechanisms;
- RoBERTa is an extension of BERT, which - avoiding the NPS task - has optimized the original model; we will run our experiments on a RoBERTa classification model.

Chapter 4

Contributions

This chapter is dedicated to the presentation of the contributions and the analysis of the obtained results. The contributions are threefold.

The first one (see section 4.1) forms the foundation of the work and concerns the presentation of an original labeled resource for SPD. The purpose of this resource is to fill the gap highlighted in the literature and to address the impossibility of finding freely available resources labeled by PDCs. The second type of contribution concerns the proposal of a model based on transformer networks with excellent results for the identification of sentences with sensitive content and the discrimination of 5 macro-PDCs (section 4.2). The model was evaluated on the aforementioned resource. The third contribution presents the model based on compositional frames and an original sub-symbolic approach to the SID task (section 4.3). Also in this case the model has been evaluated on our resource and - after a results analysis - advantages and limitations have been drawn.

4.1 SPeDaC: a Sensitive Personal Data Categories corpus

In this section, we present our first contribution to the SID task: a new resource for sensitive information and PDCs detection. In section 2.5, a crucial lack in the literature on SID concerning the presence of a benchmark has been highlighted, as the difficulty of finding an annotated corpus of sensitive personal information. As previously seen, works in literature differ greatly from each other, often starting from the investigated domain. Therefore, the first contribution of the work intends to fill this gap and proposes a resource for the SID task that represents the domain of PDCs: the Sensitive Personal Data Categories corpora (SPeDaC).

In particular, three datasets have been proposed which have two main purposes:

- The identification of sentences with sensitive personal content and sentences that in no way can be considered sensitive;
- The identification within sensitive sentences of the PDCs expressed. This identification can occur at a macro-category or a fine-grained-category level.

In section 4.1.2, the feasibility analysis that helped us to design a precise domain of investigation has been described. SPeDaC is introduced highlighting the ethical implications and disclosures to take into consideration when working on this particular field (section 4.1.1). In section 4.1.3, the structure of the resource is described, while in section 4.1.4 the methodology adopted to construct, annotate, evaluate and release the datasets has been deepened. Relevance and limits are presented in the final section 4.1.5.

4.1.1 Ethical disclosure

The automatic processing of sensitive data and the construction of a sensitive corpus imply a necessary reflection on the ethical aspects and improper uses that can derive from this type of research. In fact, due to legal and institutional concerns, when we talk about privacy and sensitive information in Natural Language Processing research, it is difficult to access relevant data. Sensitive data processing for scientific uses is regulated by GDPR in the EU in order to be consensual, fair, and transparent. The legal constraints concern (i) the explicit scientific purpose: it is to be understood in its broadest meaning and can include fundamental, applied and privately funded research [162]; (ii) a limited time window within which such data will be used; (iii) the data subject has to give explicit consent.

If we aim to offer a new benchmark for SID, this would imply first that the data could be released in the future without too strict constraints of use in terms of time. A way to reduce the sensitivity risk of data and to overcome the just mentioned problem is to work on derived data. These are types of data derived from original texts that cannot be reconstructed and - if sanitized - cannot be reconducted to identifiable individuals.

Weidinger et al. [179] talk about the ethical and social risks of harm from LMs and in particular information hazards and the importance to avoid the inferences of truly sensitive information. Sensitive information can be also inferred based on simple correlational data, and without the explicit presence of sensitive data about the individual. Obviously, the risk occurs when the individual is somehow identifiable.

Following these reflections, we can affirm how SPEDAC fully complies with ethical requirements for the following reasons:

- SPEDAC collects derived data e.g., fragments of online texts (sentences);
- The data subject can never be identified, as it is made explicit neither in the original source;
- SPEDAC aims to simulate the context of sensitive information but it does not offer sensitive content per se.

Nevertheless, LMs that aim to identify or infer SPD can directly or indirectly expose individuals to privacy risk. Even if the LMs are constructed not revealing any type of sensitive information, the model may reveal true information when applied to other contexts and used with malicious aims. For this reason, it must be ensured that the reuse of the proposed models respects the original ethical purposes. To ensure the benevolent use of the resource, control criteria are described in the release section. These criteria have been adopted for the same reasons also for the release of the KG of PD (section 4.3).

4.1.2 Feasibility Analysis of PDCs

As mentioned, this work - based on the PDCs investigation - starts from an authoritative resource such as the DPV. The resource has been used as a reference taxonomy for the definition of the PDCs of our interest; for this reason, it is necessary to underline an important aspect: the DPV receives the contributions of a very active W3C community; this means periodically updated versions of the resource with the inclusion of new PDCs. The most recently released version is version 1 of December 5th, 2022, to which a section dedicated to PDCs has been added[1]. The extension currently counts 206 different categories, but it can only be considered a continuously expanding resource.

The PDCs listed are in fact of various natures and therefore how they manifest and can be identified can be very different. We have discriminated the main criteria that distinguish the SID task and we have conducted a feasibility analysis, defining the scope of investigation even more in detail. The main criterion to be taken into consideration for the task of PD identification from text concerns the expression of the category in the form of textual data. A second criterion allows the identification of categories of greater interest over others. The categories are divided as follows:

- **Macro-categories**: in the taxonomic organization they are the high-level categories to which all the more specific PDCs belong. Their identification is therefore implicit in the identification of the nested categories. The macro-categories are six, highlighted in section 2.4.2: INTERNAL; EXTERNAL; SOCIAL; FINANCIAL; TRACKING; HISTORICAL.

- **Categories identifiable through textual analysis**: these are categories that can be frequently expressed verbatim and whose expression can be syntactically complex. They are not alphanumeric sequences or codes easily identifiable through REs, but they can be expressed in natural language depending strongly on the combination of words and the context of the sentence. 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. These are the categories that we have mainly decided to investigate, as, on the one hand, they are not conceptually too vague, and, on the other, they allow us to test the contextualization capabilities of the model.

- **Broad-boundaries categories**: these categories can be defined as characterized by (i) a high degree of vagueness; (ii) a high degree of extension and applicability; (iii) whose sensitivity classification is characterized by a high degree of ambiguity. An example is the INTENTION category. INTENTION is a sub-category of :PREFERENCE:INTERNAL and refers to '*Information about an individual's intentions*'. Any speech act could be defined as an intention or at least a communicative intention [61], and this causes difficulty in assigning sensitivity that goes beyond the due verification of (i) attribution of PD to an identifiable subject; (ii) exclusion of cases such as hypothesis or irony. These categories, due 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**: easily identifiable categories through REs and fixed sequences e.g., CREDIT CARD NUMBER, TAX CODE. Such categories (PII) have already been heavily explored in the literature for a long time. You can find tool markets offered by large companies e.g., Microsoft [9], that offer this type of identification. It seemed appropriate to

[1] As mentioned above https://w3c.github.io/dpv/dpv-pd/

focus our analysis on the most challenging and least explored categories, which could at the same time give us the possibility to analyze more complex and context-aware identification techniques.

- **Categories identifiable mainly through non-textual elements**: these depend totally 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*'.

In Table 4.1, an overview of the analysis, while in Appendix C.2 the assignments of all the PDCs are listed in detail.

N.	Type
6	Macro-categories
90	Identifiable through textual elements
26	Broad boundaries categories
30	Uniquely identifiable
54	Identifiable mainly through non-textual elements

Table 4.1: Feasibility analysis of the 206 PDCs of DPV

Analyzing, in particular, the section of the 90 categories identifiable through textual elements, most of the PDCs belong to the SPECIAL DATA, SOCIAL and EXTERNAL macro-categories. In general, the structure arrives at four levels of hierarchy (see Fig. 4.1, 4.2, 4.3, 4.4). 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.

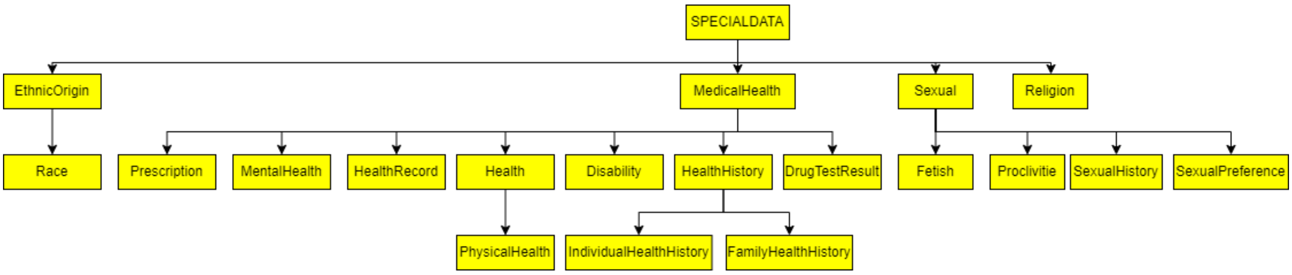


Figure 4.1: PDCs investigated :SPECIALDATA

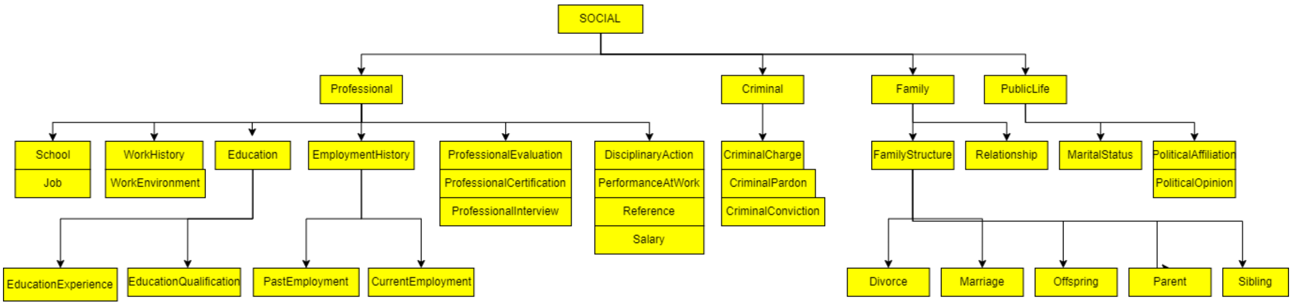


Figure 4.2: PDCs investigated :SOCIAL

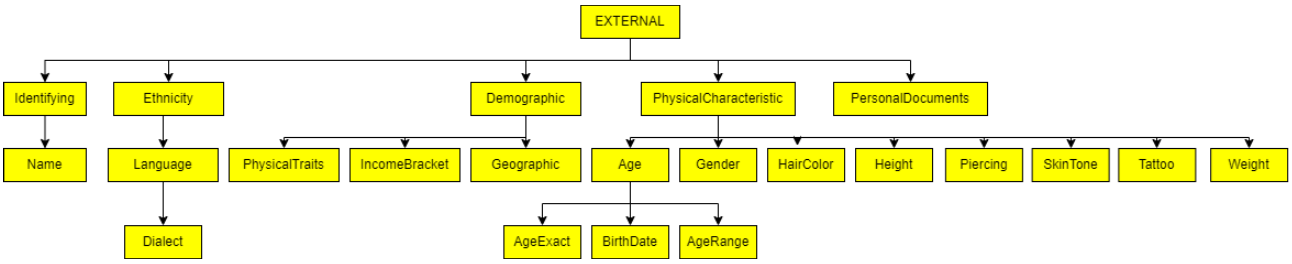


Figure 4.3: PDCs investigated :EXTERNAL

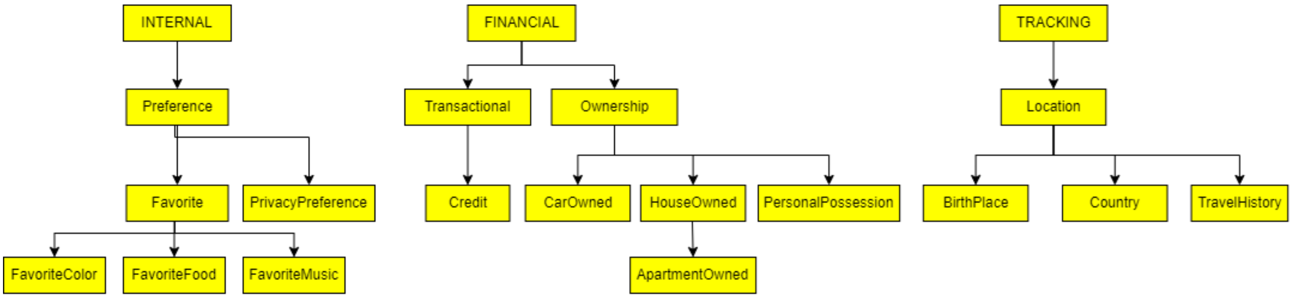


Figure 4.4: PDCs investigated :INTERNAL, :FINANCIAL & TRACKING

4.1.3 Structure

The SPeDaC resource is divided into three different corpora with different aims and labels. An adopted common methodology is described in the next section 4.1.4. The resource consists of collections of fully labeled sentences (sentence-level labeling).

SPeDaC 1 The aim of the corpus is the identification and discrimination of sensitive sentences from non-sensitive ones. The dataset counts 10,675 sentences (see Table 4.5) and has two target labels:

- **0 | NON-SENSITIVE**, to indicate sentences without sensitive content;
- **1 | SENSITIVE**, to indicate sentences with sensitive content.

The collected sentences are represented in a balanced way 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 their sensitivity. For that reason, this part of the corpus can also be called ‘adversarial corpus’. Examples are given below (see Table 4.2).

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

Table 4.2: Examples from SPeDaC 1

Other examples try to represent the penumbra cases: sentences characterized by strong ambiguity, due for example to hypothetical expressions, expressions of desire, supposition, irony,

Sentence	**Label**
I am a professional cellist with years of performing and teaching.	[SENSITIVE]
I had great hopes of being an air hostess so that I could travel to so many places then I heard about a plane crash and that kind of threw me off that idea.	[NON-SENSITIVE]
I will not let my disability define me, I am stronger than that.	[SENSITIVE]
If you have a disability, this is an absolute right.	[NON-SENSITIVE]

Table 4.3: Example of penumbra cases from SPEDAC 1

Macro PDC labels	**% labels in SPEDAC 2**
Special Data	26.42%
Financial and Tracking	12.64%
Social	30.49%
Internal	8.46%
External	21.99%

Table 4.4: % labels of macro PDC in SPEDAC 2

etc. (see Table 4.3).

SPEDAC 2 The aim of the corpus is the identification of the PDC macro-category within sensitive sentences. The dataset has 5,133 sentences (see Table 4.5) and 5 labels. The 5 target labels of the dataset 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 the PDC LIFE HISTORY, which - following our feasibility analysis - is a broad-boundaries category). The sentences retrieved to their macro-categories represent in any case the specific PDCs considered in a balanced way, i.e., we have taken into account approximately the same number of examples (100 sentences) for each fine-grained category.

The percentage of representation of the macro categories in the corpus, which depends on the number of specific categories they include, is represented in Table 4.4.

SPEDAC 3 The aim of the corpus is the identification of the fine-grained PDC within sensitive sentences. The PDCs present in SPEDAC 3 are the same considered in SPEDAC 2 plus those belonging to the HISTORICAL macro-category excluded by SPEDAC 2. The dataset counts 5,562 sentences (see Table 4.5); some of the 90 PDCs have been combined due to their similarity. A list of the labels in SPEDAC 3 can be found in Appendix D.1.

	SPeDaC 1	SPeDaC 2	SPeDaC 3
#Sentences	10,675	5,133	5,562
#Tokens	270,904	134,860	157,508

Table 4.5: Size of SPeDaC 1, SPeDaC 2 and SPeDaC 3

4.1.4 Methodology

To construct the corpus, texts from the English TenTen corpus have been collected, using SketchEngine[2] as an extraction tool.

TenTen corpus. The TenTen corpus family is a large resource, made up of texts collected from the Internet [79]. The TenTen corpora are available in more than 40 languages. The TenTen corpora are constructed as follows:

- Crawling from the Web using Spiderling tool, a linguistic-purpose tool [161];
- The sample texts were checked manually and content with poor-quality text and spam was removed;
- Text is tokenized;
- Duplicate parts are removed;
- Corpus texts are lemmatized and POS tagging.

The English version uses the Penn Treebank tagset with SketchEngine modifications[3].

All the TenTen corpora are available with the Sketch Engine [82]. The most recent version of the English TenTen corpus (enTenTen2020) consists of 36 billion words. The texts were downloaded between 2019 and 2021. They come from different domains (UK domain .uk, Australian domain .au, Canadian domain .ca, US domain .us, New Zealand domain .nz, EU domain .eu), different textual genres (news, discussion, blog, legal), and topics (reference, society, arts, technology, business, sports, science, health, home, recreation, games); the 6.8% of the corpus comes from English Wikipedia pages.

It seemed to us that the corpus could represent online language in a sufficiently complete and not too domain-specific way.

Sketch Engine. The SketchEngine is a corpus query system that, analyzing texts, can retrieve word sketches, similarities, and grammatical relations between words. In particular, it can operate at a word, sentence, or text level. We can type a word and see the most typical combinations, associate a thesaurus with synonyms and similar words, comparing collocations between words; we can type a sentence and see a WordSketch with all the words that typically appear in the same sentence, look up an example of the sentence in the context or search patterns, and look up translations; from a text, it is possible to create a word frequency list, extract keywords and terms, create a glossary of bilingual terminology, create a list frequency of multi-word expressions, discover neologisms or words not used, and see the POS tagging labels. The tool contains

[2] https://www.sketchengine.eu/ (last access January 18, 2023)

[3] https://www.sketchengine.eu/english-treetagger-pipeline-2/ (last access January 18, 2023)

600 ready-to-use corpora in more than 90 languages. Every language counts more than 60 billion words to truly represent it.

SketchEngine allowed us to easily create a subcorpus of enTenTen suitable for our purposes, thanks to the search capabilities of one or more words within different contexts and to download the textual material that the tool interface presents.

Linguistic constraints. For SPeDaC1, a dataset of adversarial sentences has been constructed. 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. To do that, linguistic constraints (i) general or recurrent, and (ii) specific or *ad hoc* for every PDC have been identified. This facilitated the process of sentences, selection, and labeling. General or recurrent linguistic constraints follow the SPD criteria highlighted in Fig. 2.3, taking into account the importance of the relationship between the PDC and the subject to which it refers, that could be identifiable. The corpus we work on does not have truly identifiable subjects, but we assume that the identifiable subject (via account, or through the device used) often corresponds to the person who writes ('I'). Therefore, the first linguistic constraints concern for example the presence of a first singular person pronoun or a first singular person possessive adjective close to the other search constraint words representing PDCs.

The specific linguistic constraints concern the keywords which could better represent every PDC and could be expressed as (i) single word, (ii) multi-word expressions, or (iii) collocations. The sentences that could represent the AGE category have recurrent keywords such as '*age*', as well as collocations and multi-word expressions such as '*[have] ... years old*'.

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. In this way, the adversarial corpus is more competitive and aims to disambiguate the sentence in relation to the subject linked to the PDC.

The adversarial corpus can represent at the same time strongly ambiguous sentences characterized by a further typology of constraints, which we call adversarial constraints. Those sentences are defined as penumbra cases (see section 2.2), in which there is a PDC, an identifiable subject, and a relationship between them, but the content cannot be considered sensitive (i) because they are simple citations, or (ii) because the wider context in which they appear cannot fit into the dimension of reality (hypothetical or ironic sentences). Adversarial constraints - regarding the first point (i) - therefore consist of citation expressions e.g., '*[he] [say]*', '*[article][say]*', '*[he][state]*' etc.

Regarding the second point (ii), they concern the dimension of unreality or supposition e.g., verbs such as '*suppose*', '*imagine*', '*guess*', '*hope*', or adverbs such as '*maybe*'. We are well aware of the complexity that characterizes the automatic identification of irony, as a creative linguistic phenomenon and therefore difficult to be defined in formal terms [142]. Semi-automatically grasping ironic sentences for the adversarial corpus presented therefore some difficulties; however, the adversarial constraints that recover some ironic sentences are expressions related to the joke and the questioning of what has just been stated e.g., '*just kidding*', '*I [be] joking*'.

Balancing and cleaning. About 100 sentences for each PDC have been collected, also paying attention to the balancing in the construction of SPeDaC 1. Indeed, here the sensitive sentences representing each PDC correspond to an equal number of adversarial sentences. For some PDCs the retrieval of 100 sensitive sentences was difficult and they are therefore less represented in the corpus. These PDCs are the following: CRIMINAL, CRIMINAL CONVICTION, CRIMINAL CHARGE, DISCIPLINARY ACTION, INCOME BRACKET, PRIVACY PREFERENCE, PROFESSIONAL EVALUATION, PROFESSIONAL INTERVIEW, SALARY, SKIN TONE.

Other PDCs, as already mentioned, have been combined. Once the corpus was built and tagged by sentence, it has been cleaned up:

1. Sentence delimiters of SketchEngine have been removed;

2. Duplicate sentences have been removed;

3. Too long sentences have been analyzed, keeping only the interesting portion.

Labeling. The corpora are sentence-level labeled. SPEDAC 1 presents binary labeling, where 0 corresponds to a non-sensitive sentence and 1 to a sensitive one. SPEDAC 2 presents a macro-labeling following the 5 macro-categories of PDCs; SPEDAC 3 a fine-grained labeling following the specific PDCs.

INCEpTION [84] has been used as annotation tool. INCEpTION is a recent platform that allows interactive and semantic annotation. A set of customized labels has been created and the resource has been released in WebAnno TSV v3.3 format.

Inter-annotator agreement (IAA). A high level of IAA highlights the goodness and reproducibility of an annotation paradigm and it's a prerequisite for demonstrating the validity of a coding scheme. The IAA can vary also in relation to the level of experience of the annotators e.g., mixed groups of experts and non-experts can decrease the IAA reliability.

To measure the goodness of our annotations, corpora subsets have been created and a group of expert linguists has been involved to annotate them. The basis given to them for annotation was the taxonomy of DPV-PD. The sentences have been randomly selected and each category is represented in a balanced way.

1. **Task 1**. SPEDAC 1: four annotators had to binary classify 100 sentences as sensitive or non-sensitive (0|1). They received the taxonomy as a reference and they were asked 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 DPV-PD;

2. **Task 2**. SPEDAC 2: three annotators had to classify 150 sentences over the 5 macro-categories of 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. **Task 3**. SPEDAC 3: since the specific PDCs are very numerous, and the effort to learn taxonomy would have been considerable, the task has been limited to the validation of our first labeling on 50 sentences. A group of four annotators contributed to it. They were asked to compare the specific PDCs with which they found the sentences labeled with the definition given in the DPV-PD.

The score agreement has been measured by aggregating the original annotation with the others, and Krippendorff's alpha (α) coefficient [69] has been used as a metric. Krippendorff's (α) is the measure frequently adopted when the annotators are three or more. The metric measures

Kappa measure	Strength of agreement
<0.0	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Table 4.6: Score agreement interpretation k [93]

the IAA in terms of disagreement and it can be expressed as follows:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where *D0* is the disagreement observed, while *D1* is the disagreement expected by chance. By not imposing a minimum number of items, it mitigates the statistical effects of low sample size datasets and ignores missing data that may be present in collaborative work. Values range from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. (α) $\geq$ .800 is usually considered a high agreement, .667 $\geq$ (α) $\geq$ .800 [89] or .610 $\geq$ (α) $\geq$ .800 [93] is an acceptable agreement. As for the reference thresholds, the various proposals of the scholars highlight their arbitrary character [47]. Table 4.6 reports the proposal by Landis et al. [93].

The Krippendorff's (α) is 0.73 for SPeDaC 1, 0.82 for SPeDaC 2, and 0.87 for SPeDaC 3. In SPeDaC 1 the sentences that reported a high rate of disagreement are mostly (i) ambiguous sentences, in which potential SPD is expressed, as well as the relationship with a subject, but appear within a non-sensitive context (e.g., an example fictitious to explain a concept); (ii) sentences in which potentially SPD appears but the subject is not uniquely identifiable (often an unspecified group of people); (iii) more specific PDCs - e.g., HOUSE OWNED - are sometimes not identified as potentially sensitive and therefore the sentence is erroneously labeled as non-sensitive.

Despite obtaining an '*almost perfect*' agreement score, SPeDaC 2 and SPeDaC 3 sometimes have multi-labeling problems in the same sentence (the problem will be discussed in particular in the next section 4.1.5).

Release. Due to the ethical disclosure described above (see section 4.1.1), the corpora have not been publicly released. A description of the available resource can be found on GitHub https://github.com/Gaia-G/SPeDaC-corpora, but the SPeDaC download is bounded to the prior signing of the user of an agreement that establishes the ethical research purposes.

Splitting. The datasets have been also used for the experimental process with DL neural networks (see section 4.2). Every dataset has been randomly divided into three parts for the experimental process: 70% training set, 10% development set, and 20% test set (see Table 4.7).

	SPEDAC 1	SPEDAC 2	SPEDAC 3
TRAINING set	7611	3695	3893
VALIDATION set	846	411	556
TEST set	2218	1027	1112

Table 4.7: Size of datasets used for experiment

4.1.5 Relevance and limits

SPEDAC was first used as dataset for the detection model; in particular, SPEDAC 1, SPEDAC 2 and SPEDAC 3 have been tested using the DL approach, and at the same time more traditional ML models. SPEDAC 3 was used in particular for the evaluation of PRIVAFRAME, the KG-based method.

The construction of the resource is relevant because it offers a training dataset and a benchmark not yet present in SID literature and in the PD domain. The advantages and particularities that SPEDAC provides are the following:

- A large and rich domain. SPEDAC is a resource that, with its 3 datasets, is configured as (i) multitask (identification of sensitive content, identification of the type of sensitive content) and (ii) multigrained (macro and fine-grained identification level). Regarding the fine-grained level, it covers a wide range of PDCs, which refer to an updated and comprehensive taxonomy. This variety also allows, if necessary, to adapt and reuse the resource according to the types of PD in which one is interested;

- The resource is already presented with a high IAA index, an evaluation of LMs and semantic models, and a baseline;

- The texts are taken from various online resources. The resource can therefore be expanded without particular difficulties in finding additional material;

- Although the resource deals with the subject of sensitive information, it cannot be considered in any way as revealing true sensitive information and this facilitates its sharing and reuse.

Despite this, some limitations should also be highlighted. The most evident detected both during the IAA process and the experimental process concerns the multi-labeling over the same sentence. It is not uncommon for a sentence to contain more than one type of sensitive data (see Fig. 4.5). This problem would not concern SPEDAC 1, therefore a single sensitive personal information makes the entire sentence sensitive and is therefore labeled with 1, whereas it concerns the category labels.

SPEDAC 2 currently has only one label per sentence. For SPEDAC 3, on the other hand, through a manual process, a multi-labeled sentence-level dataset has been realized.

The limit of single labeling will be highlighted during the classification experiments (section 4.2.4).

Other limits may relate to the broadness of the corpus, in particular:

- Quantitative aspect: the corpora could be expanded in terms of number of sentences;

- Balancing of the corpora among all the PDCs: as mentioned, it was sometimes difficult to reach the number of examples set for each PDC. The enlargement of the corpus could also focus on this aspect;

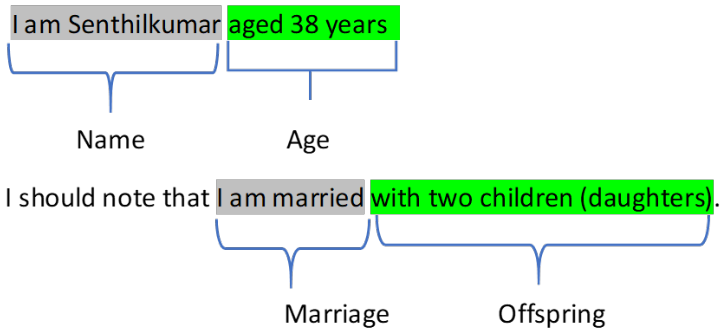

I am Senthilkumar aged 38 years
Name
Age
I should note that I am married with two children (daughters).
Marriage
Offspring

Figure 4.5: Example of multi-PDCs sentences

- Limitation to some PDCs: the resource represents the PDCs analyzed as identifiable through textual elements. Currently, for example, it excludes broad-boundaries categories, the representation of which could constitute a follow-up of the resource;
- The corpora could become multi-language and, by using the same search tools (SketchEngine), the same reference taxonomy (DPV) and the same annotation classes, could be extended to new languages.

4.2 Transformer models for SID

As we have seen in section 2.3, the most recent approaches in the literature concern the implementation of BERT-like architectures, and the results of the pre-trained models are more than satisfying.

The LMs approach is bottom-up. In this chapter, we describe how it fits into a hybrid model (section 4.2.1), all the models used for comparison (section 4.2.2), the experiments conducted to evaluate it (section 4.2.3), the results (section 4.2.4) and its advantages and disadvantages in SID (section 4.2.5).

4.2.1 Methodology

Based on the results achieved in the literature, we decided to experiment with transformer models on the PDCs domain. In particular, we have drawn two classification layers (Fig. 4.6):

1. The first layer concerns the discrimination between a sentence with sensitive content and a neutral sentence. The training set to which we refer is SPeDaC 1. It is a binary classification (y|n).
2. The second layer, which acts only on the sentences considered sensitive, concerns the identification of the macro-category of SPD. The training set is SPeDaC 2. It is a multiclass classification with 5 target labels.
3. The third layer concerns the fine-grained identification of the specific PDC. The training set is SPeDaC 3. It is a multiclass classification with 61 target labels. We will see in the next section 4.3 the advantages of carrying out this classification with a knowledge graph approach.

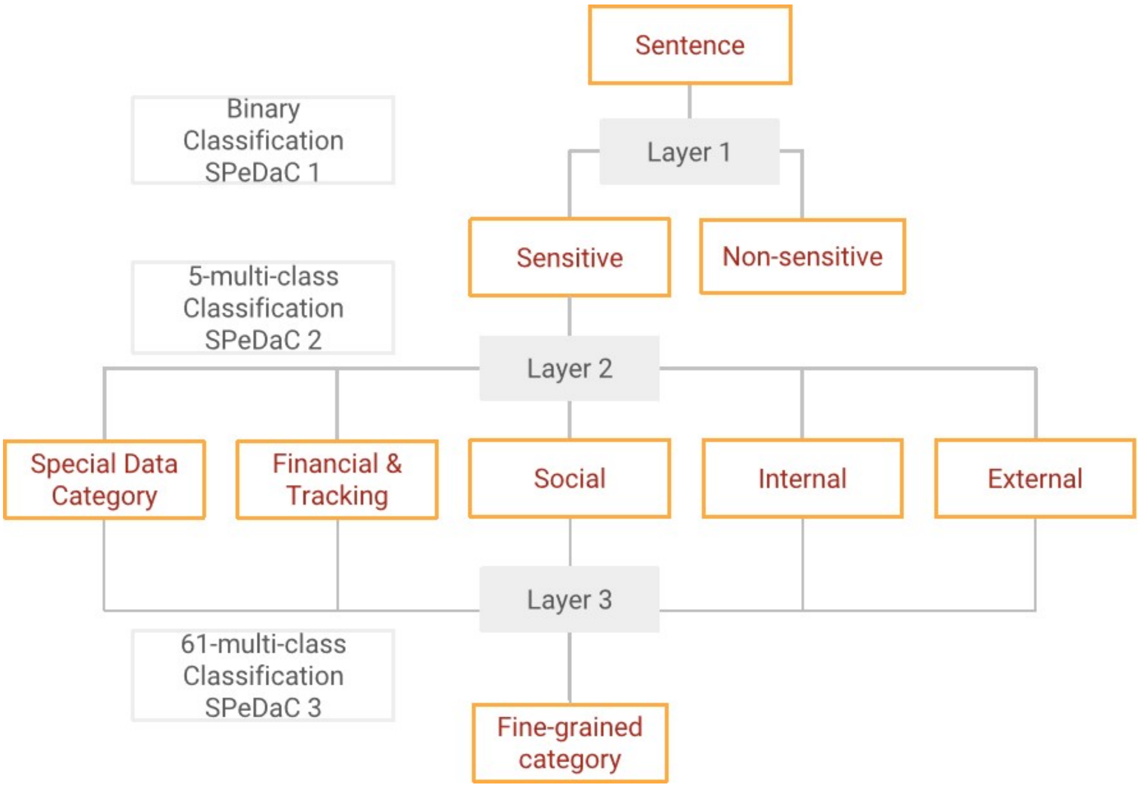


Figure 4.6: RoBERTa classification layers

The transformer model used is RoBERTa, but we considered as appropriate in the performance evaluation to test other ML models in comparison:

1. ZeroR used as the baseline
2. K-Nearest Neighbors (K-NN)
3. SVMs
4. Logistic Regression (LR)

4.2.2 Models

Here you can find a description of the ML and transformer models used in the experiments. It was included as a pure indicator of minimal rough classification results.

Zero Rate. The baseline of the models was calculated using the Zero Rate (ZeroR) classifier. This method draws up the most-frequent baseline by roughly classifying all instances as corresponding to the most-frequent class.

K-NN. K-NN is an algorithm typically used both for classification and regression, which is based on similar characteristics of neighboring features [178]. The KNN classifier is instance-based learning: it does not build a general internal model, but stores instances of the training data. An instance is classified based on a plurality vote of its closest neighbors. The data class that has the greatest number of representatives within the closest neighbors to the instance will be the predicted one. The number of neighbors to consider is a parameter of the model to be established (k). The following formula calculates the distance vector:

$$sim(d_i, d_j) = \frac{\sum_{k=0}^{n} W_{ik} \times Wjk}{\sqrt{(\sum_{k=0}^{n} W_{ik}^2)(\sum_{k=0}^{n} W_{jk}^2)}}$$

Where $sim(d_i, d_j)$ represents the similarity between the textual documents d_i and d_j; W_{ik} represents d_i characteristic weighing, while W_{jk} represents d_j. The similarity of each training sample and D is calculated; then it is calculated the weighted similarity of each category based on samples to output the final text belonging to the largest weights of the category:

$$p(D, C_j) = \sum_{i=1}^{k} sim(d_i, D) P_d(d_i, C_j)$$

$p(D, C_j)$ represents the similarity between d_i and C_j that can be established in binary terms:

$$P(d_i, C_j) = \begin{cases} 1 & d_i \text{ belongs to category } C_j \text{ samples} \\ 0 & d_i \text{ does not belong to category } C_j \text{ samples} \end{cases}$$

To resume, the inputs of the model are (i) training data, (ii) the neighboring number k, and (iii) test data.

The Big O notation [81] formally defines the time complexity of the models. The time complexity of K-NN is equal to the product of k=number of neighbors; d=number of data points; and n=number of neurons/data dimensionality. The time complexity formula of the models used can be seen in Table 4.8.

The KNeighborsClassifier sklearn model [132] has been trained[4]. 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). Especially for binary classification, the number of neighbors should be odd to avoid finding oneself in equal situations. K-NN is considered one of the simplest algorithms in ML, this is the reason for its low results.

SVMs. SVMs model is another classic algorithm [30] capable of building both binary and multiclass classifiers. SVMs model uses tagged data to define a hyperplane in which it maps training examples, trying to maximize the gap between categories. New examples are classified based on where they are mapped in space. For multiclass classification, the same principle is used after breaking down the multiclassification problem into smaller subproblems, all of which are binary classification problems. The classification can be linear or non-linear (in this case the features on the plane will be arranged non-linearly and the classifier will draw curves). The model can also be applied to regression (SVR) using methods similar to the SVMs model for classification, but returning real and continuous outputs. A function often used for both methods is the RBF kernel, which transfers the data in a larger dimensional space and then ideally draws the dividing line.

The LIBSVM [132] linear model has been used. LIBSVM [24] is an integrated software for supporting vector classification, regression, and distribution estimation and supports multiclass classification[5]. It is based on the Sequential Minimal Optimization (SMO) algorithm [135] which solves the quadratic programming (QP) optimization problem that arises during the training of

[4] The model implemented can be found here:https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (last access January 18, 2023)

[5] The model implemented can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (last access January 18, 2023)

SVMs. SMO allows handling very large training datasets because it scales between linear and quadratic in the training set size for test problems, while the traditional SVMs model scales between linear and cubic in the training set. SMO in LIBSVM-based implementation scales between

$$O(n_{features} \times n_{samples}^2)$$

and

$$O(n_{features} \times n_{samples}^3)$$

depending on data distribution. If data is very sparse $n_{features}$ is replaced by the average number of non-zero features in a sample vector. The SMO strategy consists in decomposing the problem into a set of minimal subproblems, which can then be solved analytically. The time complexity of SVMs in libsvm is generally expressed with the formula in Table 4.8 [1].

LR. LR is a regression model implemented for binary and multiclass classification activities [17]. It is also known as the sigmoid function; the sigmoid function has the shape of an S curve and measures a probability of $0 <= x <= 1$. In multiclass or multinomial LR, the labels of the dataset are > 2. The model establishes the probability to identify the value of the dependent variable by analyzing the attributes of the input and processing a weight distribution. The probability to belong to the sample is calculated for each class through a softmax function. Given k_n classes to predict, the multimodal LR is defined as follows:

$$\hat{y}(k) = 0_0^k + 0_1^k x_1 + 0_2^k x_2 + ... + 0_n^k x_n$$

And the softmax function is the following:

$$p(k) = \sigma(\hat{y}(k))_i = \frac{e^{\hat{y}(k)_i}}{\sum_{j=1}^{K} e^{\hat{y}(k)_j (k)_j}}$$

Where i=1,...,K corresponds to classes and to the probability that an example in the training set belongs to one of the classes K.

The cost function is called the cross-entropy cost function. The gradient descent measures the optimal values to minimize the cost function and predict accurately. The time complexity is a product of the data dimensionality and the number of data inputs.

The sklearn model has been used for our experiments[6].

Transformer-based Language Model: RoBERTa. The RoBERTa model is extensively described in section 3.4. The time complexity is a product between n with an exponent of 2 and d, considered per layer [175]. The RoBERTa-base model - with pre-trained weights of the model[7] and 768 hidden dimensions - has been used.

4.2.3 Experimental process

Preprocessing and feature extraction For the comparative experiments (kNN, SVMs, and LR) we have used spaCy [74], a well-known NLP library in Python that offers state-of-the-art

[6] The model implemented can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (last access January 18, 2023)

[7] The model implemented can be found here: https://huggingface.co/roberta-base (last access January 18, 2023)

Model	**Formula**
k-NN	$O(k * n * d)$
SVMs	$O(n^3)$
LR	$O(n * d)$
RoBERTa	$O(n^2 * d)$ per layer

Table 4.8: Time complexity formula of the models following Big O notation[81]

pipelines and several pre-trained models. spaCy has been used to preprocess the text in particular for the following tasks:

- Tokenization;
- Lemmatization and conversion of each token into lowercase;
- Removal of spaces and stop words.

The feature extraction has been done using the scikit learn feature extraction from text modules. The features and English language dependent (but not domain dependent) 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 [132].

For the preprocessing of the RoBERTa experiment, we used the RoBERTa tokenizer. RoBERTa tokenizer is derived from GPT-2 tokenizer and uses a byte-level Byte-Pair-Encoding. It treats spaces as parts of tokens so a word is encoded differently depending on its position in the sentence (e.g., no space if it is at the beginning).

Experiment 1 - setting up: a binary classification has been conducted using SPEDAC 1, split as indicated in section 4.1.4, Table 4.7. The dataset has been randomly split but the three sub-datasets have been the same for all the experiments conducted. The label distribution can be observed in Table 4.9. RoBERTa has been compared to the other models described in the previous paragraph.

The model parameters were set up and tuned on the SPEDAC 1 validation set as follows:

- For the K-NN model, a 3 closest neighbors (k=3) model was considered;
- For the SVMs model, we used default parameters to set up a linear kernel;
- For the LR model, default parameters have been used;
- For the RoBERTa model, 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

	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 4.9: Label distribution in SPEDAC 1

	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%

Table 4.10: Label distribution in SPEDAC 2

the AdamW optimizer [102] with a learning rate of 1*e*-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.

Experiment 2 - setting up: a multiclass classification has been conducted using SPEDAC 2, split as indicated in section 4.1.4, Table 4.7. The label distribution can be observed in Table 4.10. The models used are the just aforementioned.

The parameters of the models, set up and tuned on SPEDAC 2 validation set, are the following:

- For the K-NN model, we considered the 3 closest neighbors (k=3);
- For the SVMs 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 divides multiclass classification into a binary classification problem by class;
- For the RoBERTa model, the setting is the same as for SPEDAC 1, and likewise reports a training accuracy epoch beyond 0.90 on the validation set.

Experiment 3 - setting up: 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. 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;
- In the first study, the results obtained with the RoBERTa model previously used were very low. To improve the RoBERTa results, a category regularization with a label smoothing technique was introduced [115] and the number of epochs in training was increased to 15.

4.2.4 Results

The results are measured in terms of accuracy, which measures the percentage of the instances correctly classified:

$$\frac{vp + vn}{vp + fp + fn + vn}$$

The performance of the models for the two tasks is shown in Table 4.11.

	Experiment 1	Experiment 2	Experiment 3
Baseline	51.04%	31.93%	32.25%
KNN	68.62%	63.78%	35.30%
SVMs	93.15%	92.30%	57.59%
LR	92.60%	92.50%	75.74%
RoBERTa	**98.20%**	**94.94%**	**77.18%**

Table 4.11: Accuracy results

As can be seen, RoBERTa reports very high results compared to the other models for the binary classification task on sensitive and non-sensitive sentence identification. SPeDaC 1, as described in section 4.1.3, is made up 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 corpus often consists of contextual elements, given the occurrence of the same linguistic patterns, the RoBERTa context-aware model turns out to be the most suitable for the task. On the other hand, in the macro-category classification, where the problem of ambiguity is less evident, the results obtained with the other models are more promising. The RoBERTa model outperformed the comparison models in all cases, surpassing the most performing by 2.44%.

Even if we can't compare the results with a gold standard, we can analyze some results from similar works in the SID literature, starting in particular from the ones based on the same transformer approach (the works are detailed in section 2.3.1.3). Pablos et al. [55] obtained very high F1 results on medical sentence-level docs (97% and 95% of average score) using a BERT sequence labeling. Guo et al. [64] - working on sentence-level - obtained an F1 99.5% on a BERT Bi-LSTM Attentional Model. The domain is personal information but, as already highlighted, the PDCs are strictly limited and often correspond to REs. A BERT-For-Sequence-Classification model used in Timmer et al. [167] performs with an accuracy of 84% on the Monsanto dataset, outperforming the result achieved with the automatic paraphrase technique on the same dataset [120]. The other studies on English, concerning the domain of personal information and conducted using different methods than the BERT-like ones, report lower results than ours [12].

The results of the third experiment on SPeDaC 3 differ significantly between the models in terms of percentage accuracy and they are significantly lower, despite the adaptations of the model have improved the first performances. They offer valid results for a benchmark on SPeDaC 3.

Predicted Class

Actual Class

	RoBERTa		SVMs	
	Non-sens	**Sens**	**Non-sens**	**Sens**
Non-sens	**0.97**	0.03	**0.92**	0.08
Sens	0.01	**0.99**	0.05	**0.95**

Table 4.12: Confusion matrix Experiment 1

4.2.4.1 Results analysis

Experiment 1 Fig. 4.7 shows a t-SNE visualization [171] 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.

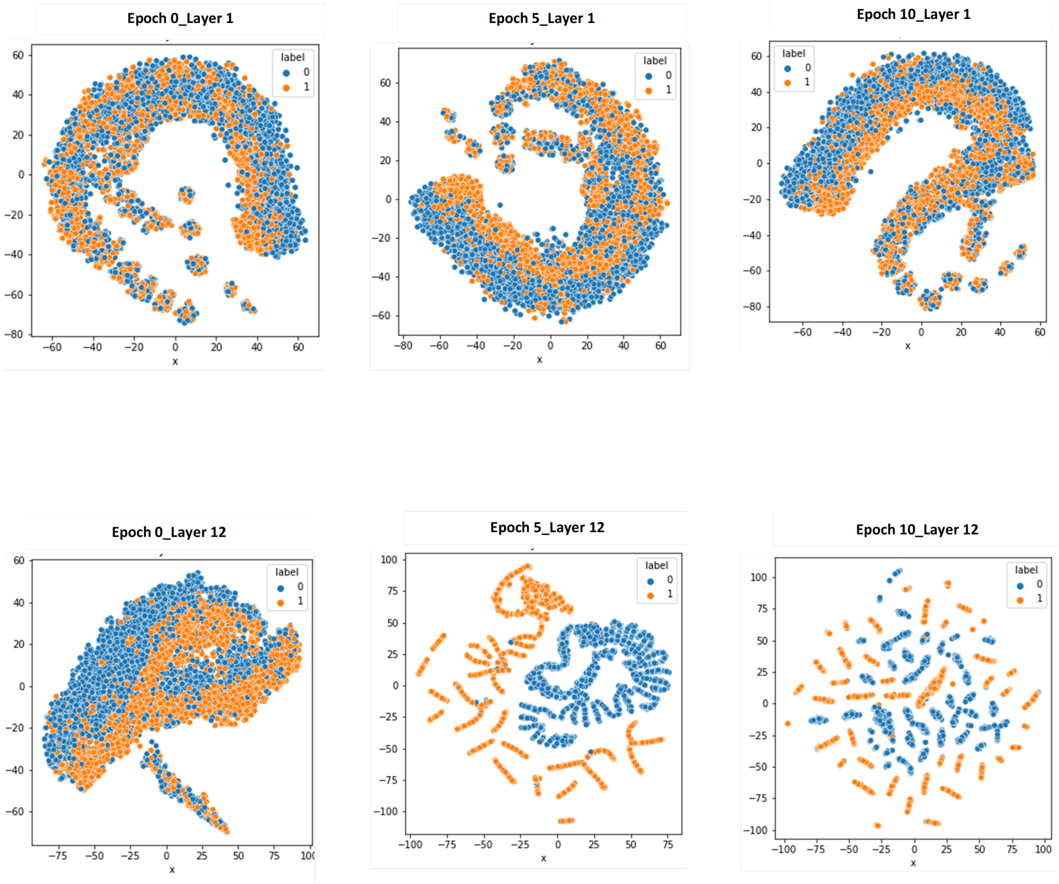


Figure 4.7: RoBERTa embeddings t-SNE visualization during fine-tuning of Experiment 1 (perplexity=30)

To better understand the behavior of the model, we report the confusion matrix that compares RoBERTa results with SVMs ones. Analyzing the errors through the confusion matrix, we see how the RoBERTa model always obtains the best performances (Table 4.12).

Both models mostly fail in identifying non-sensitive sentences, although RoBERTa is considerably more accurate. This happens even though there are more non-sensitive training sentences than sensitive ones. By analyzing errors, many sentences are misidentified as sensitive presumably due to the high rate of ambiguity they present. Table 4.13 shows some random examples of sentences labeled as sensitive, while they are not.

Examples of errors	**Actual**	**Pred**
I am almost certain I will die from a disease that I will not deserve -because nobody ever does	Non-sens	Sens
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	Non-sens	Sens
He evidently supposed I was speaking a Greek dialect, and answered in the one phrase of that tongue which he knew, and not a good phrase at that	Non-sens	Sens

Table 4.13: Analysis error Experiment 1

The first sentence refers to a disease, the second one to a hope, and the last one to a supposition. Errors are caused by the presence of expressions and keywords related to health or profession and the model in these cases is unable to discriminate assumptions or hopes useful to exclude the sensitivity of the sentence. However, it is important to note that this is not a systematic error; RoBERTa at the same time classifies as non-sensitive sentences like this hypothetical one:

'*If I receive medical care now and my insurance company pays, am I being subsidized by the insurance premiums of more-healthy people or am I being paid out of the money I 've invested in health care over lo, these many years?*'

Even if some sensitive keywords such as '*medical care*' or '*invest*' refer to a potential sensitive subject '*I*', the hypothetical construction and the interrogative tone make it not, and the model classifies it correctly. This leads us to consider that cases of ambiguity can be addressed by adding training phrases to represent them. We can consider at the same time sentences particularly ambiguous where the profession of the subject is only a guess but it could implicitly mean a real fact, e.g.:

'I'm supposed to be a movie critic, yet I keep hearing about these great new movies I've never seen.'

The sentence was labeled in the corpus as non-sensitive and thus was identified by the model. However, this is a case that highlights how useful a sensitivity scoring structure could be. This sentence could be assigned a low sensitivity score, as the condition of reality is not explicit and clear, although its potential is apparent.

Experiment 2 As for the first experiment, Fig. 4.8 shows the t-SNE visualization of the RoBERTa embeddings during the fine-tuning of training data. From this and the previous visualization, it can be seen that for both tasks, already after epoch 5, the embeddings are distinctly clustered.

As above, observing the confusion matrix of the experiment conducted on SPeDaC 2 with

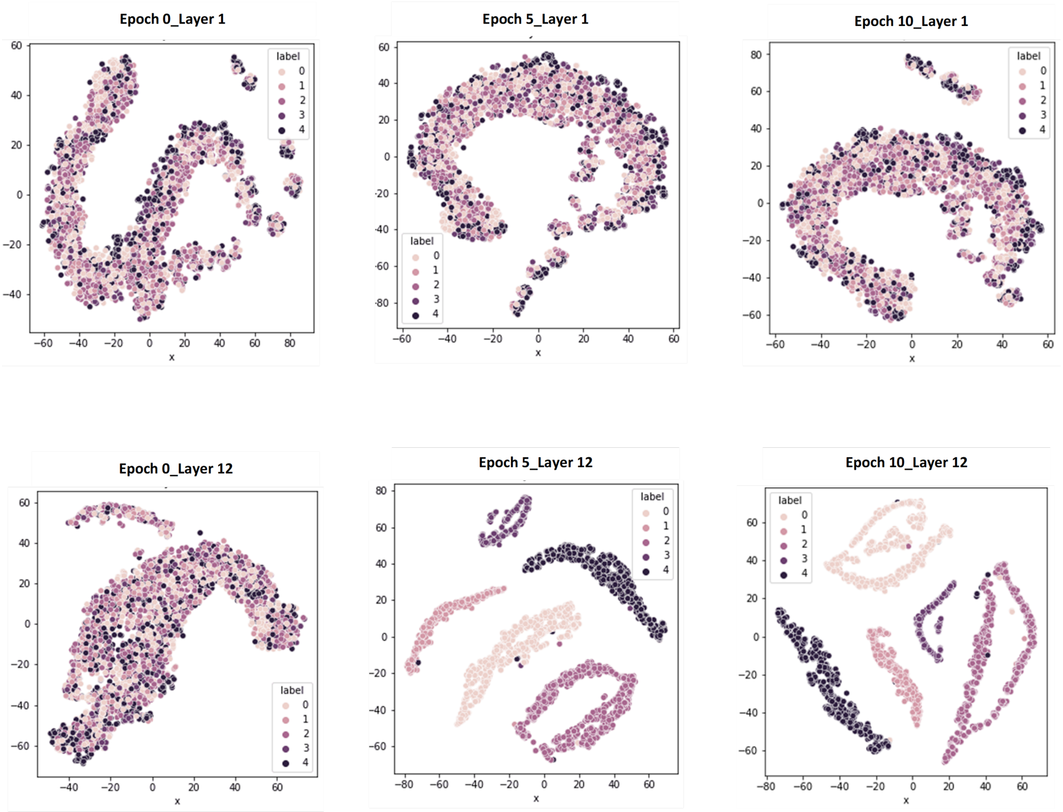


Figure 4.8: RoBERTa embeddings t-SNE visualization during fine-tuning of Experiment 2 (perplexity=30)

five target variables, we can see how RoBERTa performs always better if compared to the SVMs model (Table 4.14).

	Predicted Class									
	RoBERTa					SVMs				
Actual Class	**Spec**	**Fin**	**Soc**	**Int**	**Ext**	**Spec**	**Fin**	**Soc**	**Int**	**Ext**
Spec	**0.94**	0.00	0.04	0.00	0.02	**0.91**	0.00	0.05	0.00	0.04
Fin	0.01	**0.94**	0.05	0.00	0.00	0.01	**0.93**	0.05	0.00	0.01
Soc	0.01	0.01	**0.97**	0.01	0.00	0.02	0.02	**0.93**	0.00	0.02
Int	0.01	0.00	0.01	**0.98**	0.00	0.00	0.00	0.00	**0.98**	0.02
Ext	0.03	0.00	0.05	0.00	**0.92**	0.05	0.01	0.03	0.00	**0.91**

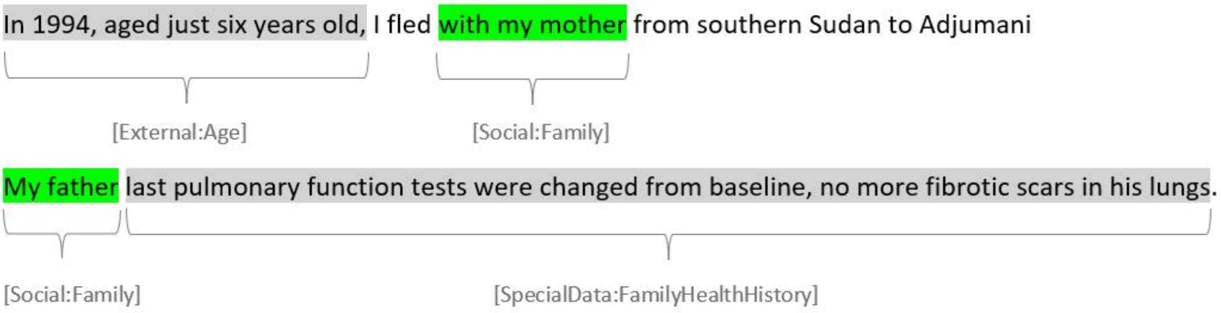
Table 4.14: Confusion matrix Experiment 2

Contrary to what might be assumed, the category with fewer training examples achieves a high accuracy score (INTERNAL). Indeed, the macro category has fewer examples but at the same time it has fewer specific PDCs that represent it (see Fig. 4.4). All the specific PDCs belonging to the macro PDC INTERNAL refer to personal preferences (FAVORITE, FAVORITE COLOR, FAVORITE MUSIC), and are therefore well identified by the model. RoBERTa mainly mistakes the macro PDC EXTERNAL for the SOCIAL or SPECIAL DATA category.

Target	Prediction	% Error RoB	% Error LR
Ethnicity	Skin Color	21.40%	21.40%
Family Health History	Drug Test Result	20.00%	20.00%
Favorite Food	Favorite	/	28.50%
Location	Country	33.30%	40.00%
Health History	Health	24.00%	/
Mental Health	Health	20.00%	20.00%
Physical Traits	Hair Color	36.80%	31.50%
Professional Evaluation	Reference	25.00%	50.00%
Reference	Employment History	20.00%	/
Reference	Professional Interview	/	20.00%
Salary	Parent	25.00%	25.00%
Salary	Credit	25.00%	/
Salary	Family Structure	/	25.00%
School	Professional Certification	31.20%	/
Sexual	Proclivitie	31.80%	/
Sexual History	Sexual	/	28.00%
Work History	Employment History	47.60%	61.90%

Table 4.15: % Errors ≥ 20% in SPEDAC 3 (RoBERTa and LR models). When '/' appears, it means a % of error < 20.

Furthermore, it can be seen that both RoBERTa and SVMs models confuse some sentences classified as SPECIAL DATA or EXTERNAL with the SOCIAL category. We can explain it by conducting a more in-depth analysis of the errors. Fig. 4.9 shows some examples: the sentences present words related to the FAMILY category (the SOCIAL macro PDC). Therefore these sentences would need a double-label in a finer granularity analysis.



In 1994, aged just six years old, I fled with my mother from southern Sudan to Adjumani

My father last pulmonary function tests were changed from baseline, no more fibrotic scars in his lungs.

Figure 4.9: Analysis error Experiment 2

Experiment 3 As for Experiment 2, the models that achieved the highest performances were the RoBERTa and the LR-based models.

By conducting an error analysis on the predictions of the two best models, we identified systematic confusion between the two labels, highlighting the errors that exceeded 20% (see Table 4.15). The confusing labels often belong to the same macro-category and present similarities in terms of keywords and linguistic patterns.

Another significant problem emerging 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 the next chapter, we will see how this problem was solved with the KG-based approach.

4.2.5 Relevance and Limits

The first great advantage of BERT-like LMs, such as RoBERTa, is the high level of accuracy they can achieve without a high effort in terms of training, fine-tuning, and optimization of parameters.

At the state-of-the-art level, if the transformer-based approach had already been tested in the literature, however, it had never been applied to PDCs classification, treated in that conceptually complex way.

The LMs also allow the processing of the analysis at a context-aware level. The model is based on sentence-level parameters to judge the presence of sensitive content or not, as well as to analyze its typology. The context-aware approach emerges particularly if we observe the results of Experiment 1, and the ability to disambiguate sentences characterized by the same linguistic constraints inserted in different contexts. It, therefore, appears to be a validly comparable approach with the most recent studies in the literature. Neerbek [120] strongly defends context-aware approaches in the SID task, proposing, as we have seen, an approach based on the automatic paraphrasing technique through RecNN. The study works on the MONSANTO dataset and categories belonging to the organizational domain and is therefore not directly comparable. However, it would be interesting to see how the RecNN model proposed by the author behaves in our domain, compared to the RoBERTa model.

If we take back the advantages and disadvantages that we had highlighted in the comparison between the top-down and bottom-up approaches (see section 3), we see how one of the characteristics of the second one concerns the starting analysis and training data work. This certainly constitutes a limitation of the transformer-based approach compared to the semantic approach. The aspect is deepened in the next paragraph.

4.3 PRIVAFRAME: a frame-based knowledge graph for SID

Since Framester covers generic knowledge, it does not necessarily cover sensitivity semantics as represented in PDCs. We have then resorted to the definition of PRIVAFRAME: a knowledge graph of new compositional frames, built on the hypothesis that each category of sensitive data can be formally described as a compositional frame. A compositional frame, as we see in chapter 3.3 is a new frame in which already existing frames and synsets are combined through logical relationships. The structure of the model and its compositional relations are described in section 4.3.1; in section 4.3.2, we point out the methodology to construct the knowledge graph and model the PDCs. PRIVAFRAME is also inserted into the hybrid architecture design. Section 4.3.3 is dedicated to the experimental process to evaluate the model; the results are reported and discussed in section 4.3.4. Relevance and limits of PRIVAFRAME (section 4.3.5) conclude the chapter.

4.3.1 Structure

PRIVAFRAME is a released resource that currently counts 86 compositional frames. The compositional frames represent not only the PDCs identifiable through text (some of them combined), but also some broad-boundaries categories. The compositional frames correspond to the PDCs listed in Appendix D.1, plus some broad-boundaries categories listed in Appendix E.1.

The resource is in turtle syntax. The PDCs are represented as follows:

```
http://www.w3.org/ns/dpv-pd#Opinion owl:equivalentClass
https://w3id.org/framester/data/framestercore/Opinion .
```

Single frame OPINION

The PDC is introduced by the DPV url and the frame by the framester url, while the relation is expressed in owl. URI schema prefixes are interpreted as follows:

```
@prefix owl: http://www.w3.org/2002/07/owl#
@prefix dpv:http://www.w3.org/ns/dpv-pd#>
@prefix frame: https://w3id.org/framester/data/framestercore/
@prefix synset: http://www.w3.org/2006/03/wn/wn30/instances/synset/
@prefix compositionalframe: https://w3id.org/framester/schema/
```

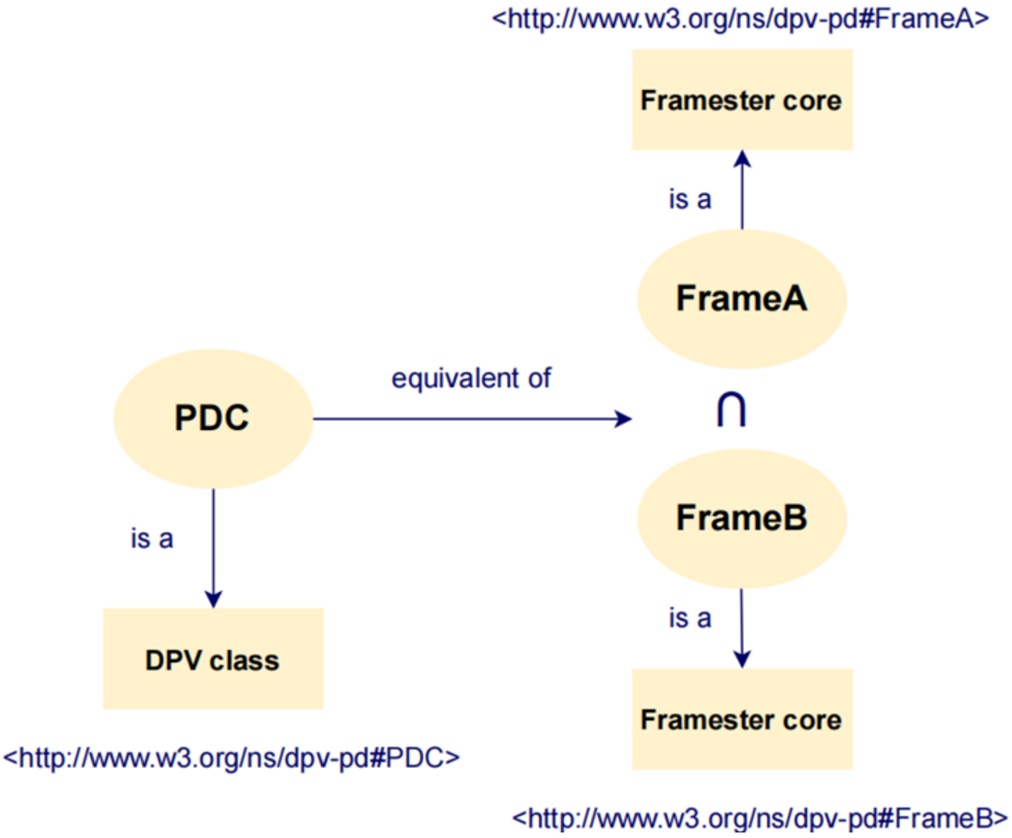

<http://www.w3.org/ns/dpv-pd#FrameA>
Framester core
is a
FrameA
equivalent of
PDC
∩
FrameB
is a
is a
DPV class
Framester core
<http://www.w3.org/ns/dpv-pd#PDC>
<http://www.w3.org/ns/dpv-pd#FrameB>

Figure 4.10: Sketch of PRIVAFRAME knowledge graph structure

As Fig. 4.10 anticipates, the relations between frames are governed by logical relations (AND/OR). The compositional relations are the following:

1. `owl:equivalentClass` equates a DPV class with a framester schema.

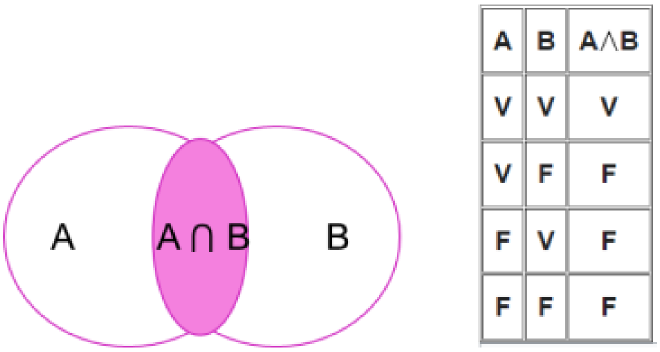
```
dpv:ReligiousBelief owl:equivalentClass fscore:ReligiousBelief
```

Single frame Religious Belief

2. `owl:intersectionOf` assumes that two or more frames coexist in the analyzed text extraction, and this coexistence is a necessary condition for the identification of a PDC (Fig. 4.11).

```
dpv:FamilyHealthHistory owl:equivalentClass [ owl:intersectionOf
(fscore:Kinship fscore:IndividualHistory fscore:MedicalConditions)
]
```

Intersectional frame Family Health History



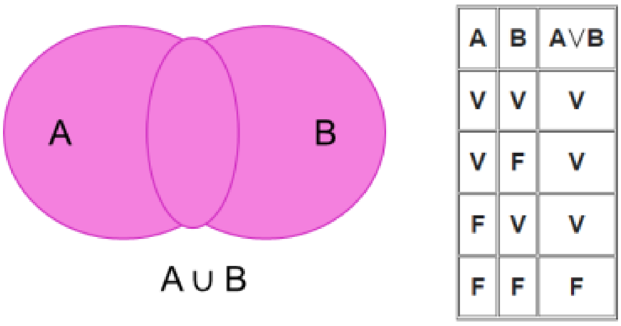
A	B	A∧B
V	V	V
V	F	F
F	V	F
F	F	F

Figure 4.11: $A \wedge B$ and table of truth

3. `owl:unionOf` assumes that at list one of the single frames composed together exists in the text and its presence allows the identification of the PDC (Fig. 4.12).

```
dpv:Job owl:equivalentClass [ owl:unionOf (fscore:Work
fscore:PeopleByVocation fscore:Being_employed) ]
```

Union frame Job



A	B	A∨B
V	V	V
V	F	V
F	V	V
F	F	F

Figure 4.12: $A \vee B$ and table of truth

4. `owl:intersectionOf (owl:unionOf)` assumes the intersection of two frames or groups of frames. For each intersectional set, the presence of only one frame of a group is enough to make the intersection valid.

```
dpv:DisciplinaryAction owl:equivalentClass [ owl:intersectionOf
( [ owl:unionOf (fscore:PeopleByVocation fscore:Work
fscore:Being_employed) ] fscore:RewardsAndPunishments) ] .
```

Compositional frame AND[OR] Disciplinary Action

5. `owl:unionOf (owl:intersectionOf)` assumes the presence of at least one group of intersected frames.

```
dpv:Age> owl:equivalentClass [ owl:unionOf (fscore:Age
fscore:PeopleByAge fscore:Aging) [ owl:intersectionOf
(fscore:People> fscore:MeasureDuration ) ] .
```

Compositional frame OR[AND] Age

Table 4.16 shows the number of compositional relationships of the resource.

Compositional Relation	**N.**
owl:equivalentClass	12
owl:intersectionOf	16
owl:unionOf	15
owl:intersectionOf (owl:unionOf)	19
owl:unionOf (owl:intersectionOf)	24

Table 4.16: Compositional relations in privaframe

PDCs can be represented not only by frames but also by synsets. Synsets mainly refer to the Wordnet resource (see section 3.3). Synsets allow the represention of particularly specific categories and create compositional frames characterized by a deeper level of detail.

The synset subalignment presents the same logical relationships. Some examples are shown below:

```
(1)
dpv:Tattoo> owl:equivalentClass synset:tattoo-noun-2  .
(2)
dpv:School> owl:equivalentClass  [ owl:unionOf
(fscore:EducationTeaching synset:diploma-noun-1
synset:degree-noun-1) ] .
(3)
dpv:Divorce> owl:equivalentClass [ owl:unionOf ( [
owl:intersectionOf (synset:marriage-noun-2 fscore:
BecomingSeparated) ] synset:divorce-noun-1 ) ] .
```

Compositional frames with synsets (1) Tattoo (2) School (3) Divorce

The Knowledge Graph has been uploaded to the Framester SPARQL endpoint[8] and it is available as Graph IRI at https://w3id.org/framester/dpv2fn.

The Framester knowledge graph can be explored through SPARQL queries:

```
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT * {

GRAPH <https://w3id.org/framester/dpv2fn> {
?s a owl:Class .
?s (owl:equivalentClass|owl:intersectionOf|owl:unionOf|rdf:rest|rdf:first)
+/rdf:first ?o
FILTER(isIRI(?o))

 }
}
limit 100
```

Query to retrieve all the alignments of privaframe graph

4.3.2 Methodology

privaframe **modeling.** A top-down approach has been adopted in the privaframe construction based on the following steps:

1. Analysis of the definition of the PDC provided by the DPV.
   E.g .: MARITAL STATUS is defined in DPV as '*Information about marital status and history*'.

2. Search for frames that can represent it conceptually.
   This research often took place in a top-down manner, starting from the exploration of

[8] It is accessible here http://etna.istc.cnr.it/framester2/sparql

generic frame cores with relevance to the PDC and then exploring the LUs associated with them to find some possible detailed representation.
E.g ., marital status is, first of all, a legal act and implies the presence of two people. The marital status could also be implicitly inferred from the act that precedes it, namely marriage. However, another PDC specification is reserved for MARRIAGE. We, therefore, identify two keywords that refer to the legal status of the spouse after the union in marriage: '*husband*' and/or '*wife*'.

The modeling of this category is quite detailed and that is why we use synsets. We linked the concept to two synsets of Wordnet: `synset:husband-noun-1` and `synset:wife-noun-1`.

The subject should then express a personal connection with '*husband*' or '*wife*' of the sentence. We, therefore, use two generic frame cores that model the concept of kinship:

`fscore:PersonalRelationship` and `fscore:Kinship`.

3. Combination of frames establishing logical relationships.
   E.g. in the case of MARITAL STATUS PDC, the relationship between the two synsets expresses a union and it is sufficient that one of the two elements is present; `synset:husband` or `synset:wife` must however be present with at least one frame that evokes a personal relationship: `fscore:PersonalRelationship` or
   `fscore:Kinship`.

   The compositional frame will therefore look like this:

```
dpv:MaritalStatus> owl:  equivalentClass [ owl:  intersectionOf
([unionOf (fscore:PersonalRelationship fscore:Kinship) ([unionOf
(synset:husband-noun -1 synset:wife-noun-1))].
```

Compositional frame MARITAL STATUS

In this way, the first modeling of the resource was carried out. Finally, the top-down approach was combined with a bottom-up approach. For this approach, we used the automatic frame extraction technique, detailed in the next paragraph 4.3.3. Some sentences containing the PDCs to be represented have been analyzed as examples and the observation of the frames identified has allowed us to bring out more quickly critical issues, possible expansions, or necessary changes to the resource. PRIVAFRAME was then systematically evaluated.

High-level methodology. PRIVAFRAME, due to its knowledge-grounding, is potentially a resource that can operate on the fine-grained identification of information. However, compositional frames evoke situations present in the text, but they are not able by themselves to define, for example, syntactic relationships between sensitive information and context. For this reason, the knowledge graph is part of a more complete model which can at the same time exploit the potential and powerful accuracy of the transformer approach shown above.

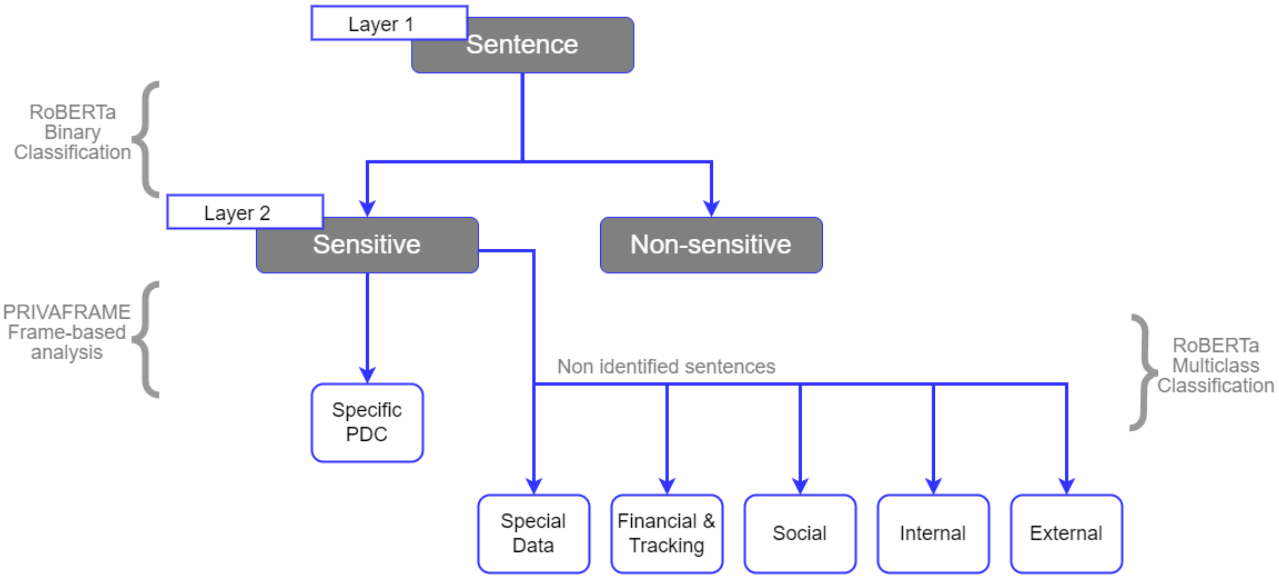


Figure 4.13: Hybrid model combining RoBERTa and PRIVAFRAME

PRIVAFRAME is therefore proposed as a second layer, which attempts to identify the specific category on text predicted as sensitive, while the multiclass classification is presented as a sheltering layer, with a more raw identification when PRIVAFRAME fails.

Like the transformer networks, the knowledge graph was evaluated in terms of performance. We illustrate the experiments in the next paragraphs.

4.3.3 Experimental process

The evaluation experiments on PRIVAFRAME have been conducted using SPeDAC 3 (see section 4.1.3). The original resource presents only one labeled PDC per sentence. The fine-grained labeled dataset has been used in this way:

- The 34% of the dataset has been used for preliminary tests to refine the model during its design and to better identify linguistic patterns and improve the compositional frames modeling;
- The rest of SPeDAC 3 constitutes the test set, which counts 3671 sentences;
- The test set has been multi-tagged: each sentence has been enriched with more than one sentence-level specific PDC;
- Some PDCs labels have been merged by conceptual similarity e.g., CRIMINAL CHARGE, CONVICTION, and PARDON, which have been considered under the more generic CRIMINAL PDC. The target labels are a total of 33. The detailed distribution can be seen in Table 4.18.
- The broad-boundaries PDCs are described as compositional frames in the knowledge graph, but not in SPeDAC. For this reason, they have been excluded from evaluation for the moment.

To analyze the sentences and automatically extract the frames, FRED [51] has been used (see section 3.2). FRED is developed in Python and it is available as a REST service. It can automatically extract Framester alignments.

The Framester alignments retrieved through FRED (framester + synsets) are compared to the PRIVAFRAME compositional frame to identify the presence of one or more PDCs.

Let's follow this example:

'I have a **civil engineer diploma** [PROFESSIONAL CERTIFICATION] of three years'

When we submit this sentence to FRED to be analyzed, we obtain an extraction as Fig. 4.14, where `fs:Documents` and `fs:PeopleByVocation` are identified.

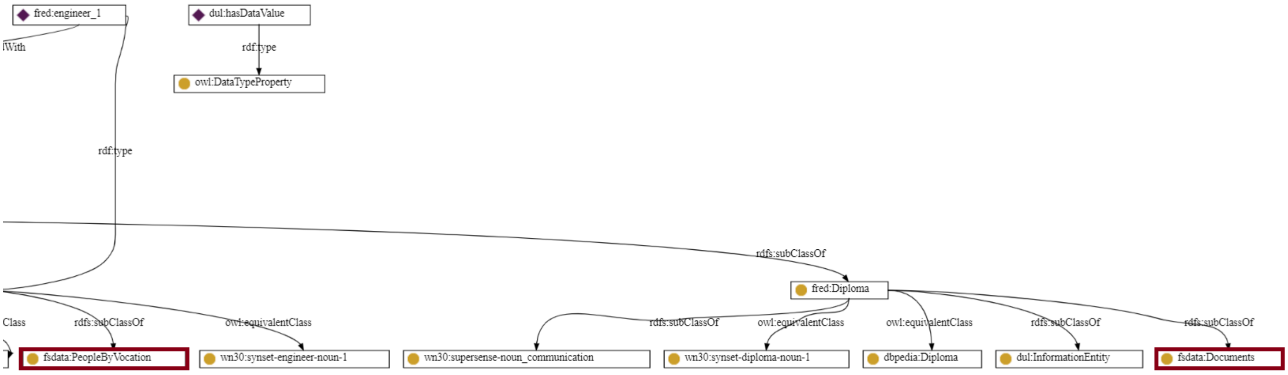


Figure 4.14: FRED extraction of *'I have a civil engineer diploma of three years'*

The two extracted frames identified intersectional frames in the PROFESSIONAL CERTIFICATION representation; it means that the specific PDC could be predicted in this case.

We developed a Python script to launch to compare the alignment of PRIVAFRAME with the FRED extraction.

The PRIVAFRAME experimental model (py.code) and the test set could be released but, as for SPEDAC (section 4.1.1), we have to consider the ethical disclosure. Indeed, the experimental process can be replicated for malicious uses. To avoid the possibility of allocating our model to uses that are opposed to our purpose of personal information protection, we have again decided to bound the download of the model to the previous signing of an ethical agreement[9].

4.3.4 Results

To compare them with the transformer approach, the results are measured in terms of accuracy. Concerning correctly identified labels, even on multi-labeled sentences, the model achieves an accuracy of 78%. 75% of the sentences (single and multi-labeled) obtain a complete identification of the PDCs labels, and 10.2% obtain partial correctness (e.g., not all the labels of the sentence have been predicted).

Let's look indeed the fine-grained predictions: the detailed number of correct predictions over actual labels can be explored in Table 4.18. It can be noted how there are some PDCs that receive a very high prediction result e.g., DISABILITY, NAME, PERSONAL POSSESSION, RELATIONSHIP; as well as we can observe particularly critical categories e.g., POLITICAL AFFILIATION, PROFESSIONAL CERTIFICATION, PROFESSIONAL EVALUATION, REFERENCE. In Table 4.17, the results are classified according to their performances, while the distribution of predicted PDCs is shown in Fig. 4.15.

[9] A general resource description and the ethical agreement can be found in the repository https://github.com/Gaia-G/PRIVAFRAME

Performance	PDCs
Excellent (+**90%**)	CRIMINAL; DISABILITY; HAIR COLOR; INCOME BRACKET; NAME; PERSONAL POSSESSION; PRESCRIPTION & DRUG RESULTS; RELATIONSHIP, DIVORCE, MARRIAGE & MARITAL STATUS; SCHOOL; SKIN TONE
Very Good (+**75%**)	HEALTH; FAVORITE; DEMOGRAPHIC, COUNTRY & LOCATION; JOB, PROFESSIONAL & EMPLOYMENT HISTORY; LANGUAGE; OFFSPRING
Good (+**65%**)	FAMILY, PARENT & SIBLING; FETISH; ETHNICITY; PROFESSIONAL INTERVIEW; RELIGION
Sufficient (+**55%**)	CREDIT & SALARY; GENDER; PHYSICAL HEALTH; SEXUAL
Critical (**-55%**)	AGE; PHYSICAL TRAITS; POLITICAL AFFILIATION; PRIVACY PREFERENCE; PROFESSIONAL CERTIFICATION; PROFESSIONAL EVALUATION; RACE; REFERENCE

Table 4.17: Score range of PDCs identification

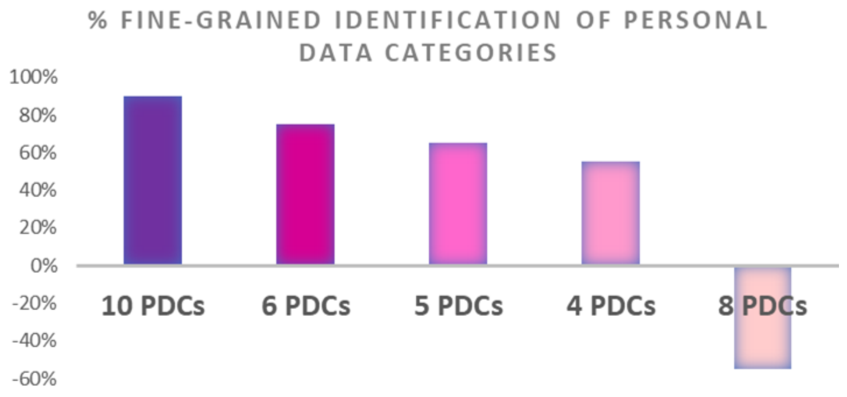


Figure 4.15: PDCs PRIVAFRAME classification

Table 4.18 also shows the number of false positives (FP). If precision considers:

$$\frac{vp}{vp + fp}$$

the model reaches 60% of the result.

It is also interesting to observe the performance achieved by the model in identifying the macro-categories. PRIVAFRAME is a multi-label model; however, in this case, we followed and measured the performance on single-category labeling, as for SPeDAC 2. PRIVAFRAME reaches 85% accuracy. In particular, the internal macro-category is not often identified and is confused with the SOCIAL class; as well as EXTERNAL, which in 10% of cases is identified as FINANCIAL & TRACKING (see Table 4.19).

PDC	**Labels**	**TP [%]**	**FP**
Age	109	57 [*52%*]	156
Credit & Salary	122	87 [*71%*]	181
Criminal (Charge, Conviction, Pardon & Offense)	16	16 [*100%*]	20
Disability	93	88 [*95%*]	2
Ethnicity	68	46 [*68%*]	16
Family, Sibling & Parent	676	488 [*72%*]	120
Favorite (Food, Color & Music)	213	171 [*80%*]	11
Fetish	51	37 [*72%*]	0
Gender	80	45 [*56%*]	4
Demographic, Country & Location	179	146 [*82%*]	376
Hair Color	80	79 [*99%*]	29
Health (Medical & Mental), Health History & Family Health History	407	308 [*76%*]	119
Income Bracket	40	39 [*97%*]	1
Job, Professional & Employment History	360	318 [*88%*]	241
Language	155	125 [*81%*]	48
Name	101	92 [*91%*]	63
Offspring	159	124 [*78%*]	16
Personal Possession, Apartment Owned, Car Owned & House Owned	300	284 [*95%*]	746
Physical Traits (Height, Weight, Piercing & Tattoo)	223	121 [*54%*]	33
Physical Health	46	28 [*61%*]	77
Political Affiliation	9	0 [*0%*]	0
Prescription & Drug Test Result	205	203 [*99%*]	1
Privacy Preference	8	1 [*12%*]	1
Professional Certification	22	0 [*0%*]	0
Professional Evaluation	5	0 [*0%*]	0
Professional Interview	42	28 [*67%*]	26
Race	54	21 [*39%*]	3
Reference	12	0 [*0%*]	0
Relationship, Divorce, Marriage & Marital Status	460	429 [*93%*]	129
Religion	69	45 [*65%*]	8
School	166	149 [*90%*]	38
Sexual, Sexual Preference, Sexual History & Proclivity	221	124 [*56%*]	3
Skin Tone	58	57 [*98%*]	6

Table 4.18: Test-set: number of PDCs labels and number of labels detected

		Predicted Class				
		Spec	**Fin**	**Soc**	**Int**	**Ext**
Actual Class	**Spec**	**0.86**	0.04	0.06	0.01	0.03
	Fin	0.00	**0.95**	0.02	0.00	0.03
	Soc	0.00	0.06	**0.88**	0.00	0.05
	Int	0.02	0.06	0.14	**0.71**	0.06
	Ext	0.02	0.10	0.07	0.01	**0.80**

Table 4.19: Confusion matrix PRIVAFRAME macro-categories

4.3.4.1 Results analysis

One of the great advantages of the PRIVAFRAME rule-based model is its explainability. It can be clearly observed which are the extracted frames and consequently the proposed assumptions of the model. In order to improve the performance, we present a detailed error analysis.

Three main types of recurring hypothesized errors can be identified, depending on different causes:

1. FRED's failure on frame extraction;
2. Lacks in compositional frame modeling;
3. Complexity in the structure of the sample sentences to be identified.

Some errors may likewise be due to dataset labeling errors, but these cannot be assumed as recurring for specific categories. The PDCs evidently critical (-55% of TP sentences) can be first of all observed. For each point analyzed, we will report in brackets the type(s) of error hypothesized:

1. AGE (*a,c*): AGE labeling is often not correctly defined. Analyzing in detail, sentences with similar structures seem not always to be identified by the model:

   '[...] i am a 31 year old woman.'
   'Hi My name is Megisiana (Megi) I am 13 years old [...]'
   'I am 24 male'

   The first sentence is correctly identified, while in the second and third sentences, the AGE label is missing. Since the linguistic patterns that elicit age information are the same in the first two sentences, the hypothesis is that this error is due to a failure to identify the frames during the FRED extraction process. As shown in the third sentence, otherwise the problem is due to the structure of the sentence, which does not contain sufficient elements for identification.

2. PHYSICAL TRAITS (*b*): the generic category includes specific PDCs, namely HEIGHT, WEIGHT, TATTOO, and PIERCING. If the PDC HEIGHT is often identified, this not happens for WEIGHT. There are no significant complexities concerning the variety or structure of the sentences, e.g.:

'[...] my weight usually ranges between 125-130 lbs'

The compositional frame is very articulated, with both AND and OR relationships:

```
dpv:Weight owl:equivalentClass [ owl:intersectionOf ([unionOf
(fscore:People fscore:Measure_mass fscore:  Dimension ) ]
syn:Weight ) ] .
```

Compositional frame WEIGHT

FRED identifies some of the frames but rarely manages to reconstruct the complete composition. A more generic rule could be modeled, losing a few points in precision. As for the TATTOO and PIERCING PDCs, the problem lies in the fact that it has not been possible to find compositional frames to adequately represent them. However, these categories are not even identified on a more generic level, such as PHYSICAL TRAITS. Again, a different modeling strategy would need to be found.

3. POLITICAL AFFILIATION, PRIVACY PREFERENCE, PROFESSIONAL CERTIFICATION, PROFESSIONAL EVALUATION, RACE and REFERENCE: these are PDCs represented by rather articulated compositional frames.

   POLITICAL AFFILIATION:

```
dpv:  PoliticalAffiliation owl:equivalentClass [
owl:intersectionOf  ( synset:party [ owl:unionOf (synset:
affiliation synset; affiliated fscore:  TakingSide) ]] .
```

Compositional frame POLITICAL AFFILIATION

PRIVACY PREFERENCE:

```
dpv:PrivacyPreference owl:equivalentClass  [owl:unionOf
([owl:intersectionOf (fscore:  Preference fscore:  Secrecy_status )
(fscore:  Secrecy_status synset:  prefer) (synset:  privacy synset:
prefer) (synset:  privacy fscore:  Preference) ] ) ] .
```

Compositional frame PRIVACY PREFERENCE

PROFESSIONAL CERTIFICATION:

```
dpv:ProfessionalCertification owl:equivalentClass [
owl:intersectionOf ( [ owl:unionOf (fscore:  PeopleByVocation
fscore:  Work fscore:  Being_employed) ] fscore:  Documents) ] .
```

Compositional frame PROFESSIONAL CERTIFICATION

PROFESSIONAL EVALUATION:

```
dpv:  ProfessionalEvaluation owl:equivalentClass [
owl:intersectionOf ( [ owl:unionOf (fscore:  PeopleByVocation
fscore:  Work fscore:  EducationTeaching fscore:  Being_employed) ]
fscore:  Assessing ) ] .
```

Compositional frame PROFESSIONAL EVALUATION

RACE:

```
dpv:Race owl:equivalentClass [ owl:unionOf ( [owl:intersectionOf
(fscore:  PeopleByOrigin fscore:  Type) ] synset:  Race ) ] .
```

Compositional frame RACE

REFERENCE:

```
dpv:  Reference owl:equivalentClass [ owl:intersectionOf ( [
owl:unionOf (fscore:  PeopleByVocation fscore:  Work fscore:
Being_employed ) ] fscore:  Attempt_suasion ) ] .
```

Compositional frame REFERENCE

As can be seen, they have a very low or zero number of FP. In particular, sentences that represent PROFESSIONAL CERTIFICATION - e.g., *'I had a diploma'* - often present double labeling (SCHOOL and PERSONAL POSSESSION).
PROFESSIONAL EVALUATION is often confused with PROFESSIONAL INTERVIEW, which is represented as follows:

```
dpv:  ProfessionalInterview owl:equivalentClass
[ owl:intersectionOf ( [owl:unionOf ( fscore:  PeopleByVocation
fscore:  Work fscore:  EducationTeaching fscore:  Being_employed )
] fscore:
Assessing synset:  article synset:  Interview ) ] .
```

Compositional frame PROFESSIONAL INTERVIEW

The sentences representing REFERENCE often are labeled with SCHOOL and PROFESSIONAL.

In all these cases, more generic modeling, or a merging of specific PDCs into a single one could address the problem. It is also advisable to increase the number of sample sentences. Those tested are structurally complex and very varied from each other.

Some PDCs are quite well identified, achieving sufficient or good results (+55% and + 65% acc.):

GENDER and RELIGION errors seem to be related to FRED's missed frame extraction. In fact, as for AGE, sentences with recurring structures are identified differently:

‘i am a 32 year old male’ (1)

‘I am a male Tanzanian aged 51 years’ (2)

‘I am Jewish, and living in the United States I have always been able to openly practice my religion’ (3)

‘So I prayed to Jesus (I am Jewish by the way), to take my asthma away [...]’ (4)

The first sentence is labeled with GENDER label, while the second is not. The third sentence is labeled with RELIGION, while the fourth is not. Other PDCs e.g., ETHNICITY, FAMILY, PARENT, SIBLING, PHYSICAL HEALTH, and SEXUAL, could perhaps improve their accuracy through an expansion of the LUs with which they are represented.

Others e.g., CREDIT & SALARY, FETISH, PROFESSIONAL INTERVIEW, may appear in a form of structurally more varied sentences and are represented by more complex and sometimes too articulated compositional frames, so first it would be advisable to intervene in modeling, but they could also be deepened with other test sentences.

Concerning the precision score, some PDCs produce a large number of FP compared to the number of TP e.g., AGE, CREDIT & SALARY, DEMOGRAPHIC, COUNTRY & LOCATION, JOB, PROFESSIONAL, EMPLOYMENT & WORK HISTORY, and PERSONAL POSSESSION. Errors can typically come from PDCs that are too large, producing higher numbers of FPs than specific ones, or from compositional frame modeling, which may need additional rules. In particular, the following observations are highlighted:

1. AGE. The sentences in which AGE is present as FP contain elements related to age not directly attributable to the subject. Age could refer to non-animated things (e.g., the car purchased by the subject) or events or subjects not directly identifiable e.g.:

   ‘My Mum had bowel cancer about 7 years ago’

   or age information that does not lead to the subject's age assumption, despite being part of his or her personal history e.g.:

   ‘From age 22 to 23, I lived in England .’

2. CREDIT & SALARY.

```
dpv:  Credit owl:equivalentClass [ owl:unionOf ( fscore:
EarningAndLosses ) [ owl:intersectionOf ( fscore:  Money fscore:
People fscore:  EarningAndLosses ) ] ) ] .
```

Compositional frame CREDIT & SALARY

The EARNINGS & LOSSES frame in the PDC compositional frame tends to expand its labeling to sentences that contain LUs attributable to gain and loss in a broad sense. E.g.:

'Twenty-two years after my divorce , I have gained a different perspective.'

3. DEMOGRAPHIC, COUNTRY & LOCATION. FP often concerns sentences that present personal information about the individual's history (WORK EMPLOYMENT, HEALTH HISTORY), in which some information concerning the individual's movements are presumed; or again, information belonging to the CAR OWNED or HOUSE OWNED PDCs in which, in the same way, movements or transfers are mentioned; e.g.:

   'I bought a home and after 6 years of living there I rented it to my first tenant.'
   'I trained and worked as an electrician for six years before deciding to go to college.'

4. JOB, PROFESSIONAL, EMPLOYMENT & WORK HISTORY. Many identifications are confused with the FAMILY and RELATIONSHIP PDCs presence, as the profession of a family member or of a person, with whom the interested subject has a relationship, is made explicit, e.g.:

   'My son has a very unusual job, he is an antique weapons restorer.'

   It is often confused with the PRESCRIPTION & DRUG RESULTS PDC, because the figure of the attending physician is appointed. In addition, many sentences are labeled with the PDC SCHOOL which also has the PROFESSIONAL label. The confusion could be reduced by introducing more specific rules that represent the PDC.

5. PERSONAL POSSESSION, OWNERSHIP, HOUSE OWNED, APARTMENT OWNED & CAR OWNED. These categories have the highest number of FP. PERSONAL POSSESSION and OWNERSHIP are very generic PDCs; it is sufficient that in the sentence the subject refers to something that belongs to him, not necessarily material to be identified. E.g.:

   'I have a terrible headache'

   If we observe the more specific CAR OWNED, HOUSE OWNED, APARTMENT OWNED, the FP is significantly reduced to 33.

For 5, and in part for 3, the problem, therefore, lies in the potential extension of PDCs; certainly, the identification becomes more precise when it is reduced to more detailed sub-PDCs. Problems 1, 2, and 4 should instead be faced with the design of additional rules that strengthen the labeling (presumably consequently finding an accuracy decrease).

4.3.5 Relevance and Limits

One of the strongest advantages of PRIVAFRAME concerns the amount of training data required: the top-down model starts from the representation of the theoretical concepts of the PDCs and does not need any training data.

On the contrary, as we have been able to observe, the transformer-based model, although it can be adapted with a fairly limited amount of data, requires training and labeled examples. The sub-symbolic model is independent of the number of categories to identify, the second does not. This makes PRIVAFRAME potentially and easily able to be expanded using only test datasets for its improvement and evaluation.

If we go back to the works on SID from the literature, it is easy to see how rule-based or ontology-based works have not recently been adopted to the detriment of neural network-based approaches. The compositional frame approach of PRIVAFRAME is a novelty. The model also allows you to investigate by sentence, therefore neither at a too general level (document-level) nor at a too specific level (word-level) which are not context-aware.

In addition to being context-aware, unlike transformer-based models, PRIVAFRAME is able to identify the segment(s) of the offending sentence(s), thanks to the labeling for frame of the analyzed sentence.

Finally, as highlighted, the explainability of the model is very high and this allows for intervention in the modeling of the categories with a high degree of precision.

We have discussed the most critical PDCs, but let's now try to observe the PDCs that have achieved high identification performance (> 90%) and low FP rate (≤10%):

- DISABILITY.

```
dpv:Disability owl:equivalentClass [ owl:unionOf (
[owl:intersectionOf (fscore:Capability fscore:MedicalConditions)
synset:  disability-noun-1 ] ) ] .
```

Compositional frame DISABILITY

- INCOME BRACKET.

```
dpv:IncomeBracket owl:equivalentClass [ owl:unionOf
(synset:income-noun-1 [ owl:intersectionOf (synset:bracket-noun-4
fscore:EarningsAndLosses) (synset:income-noun-1
synset:bracket-noun-4 ) ] ) ] .
```

Compositional frame INCOME BRACKET

- PRESCRIPTION & DRUG RESULTS.

```
dpv:Prescription owl:equivalentClass [ owl:unionOf ( [
owl:intersectionOf (fscore:  MedicalConditions fscore:Cure
fscore:  Medical_intervention) (fscore:Intoxicants
fscore:MedicalProfessionals ) ] ) ] .

dpv:DrugTestResult owl:equivalentClass [ owl:intersectionOf
(fscore:  Intoxicants fscore:Addiction) ] .
```

Compositional frame PRESCRIPTION & DRUG RESULTS

- SKIN TONE.

```
dpv:  SkinTone owl:equivalentClass [ owl:unionOf ( [
owl:intersectionOf (fscore:Color synset:skin-noun-1 )
(fscore:Possession synset:skin-noun-1 ) ] ) ] .
```

Compositional frame SKIN TONE

The specificity of the categories, compared to more generic categories such as HEALTH, CREDIT or PHYSICAL TRAITS, allows them to be identified with greater precision. Surely the identification accuracy also derives from the variety of example sentences: the sentences that represent these more specific categories are characterized by less variety and more frequency of recurring linguistic patterns.

Once again, the sub-symbolic model compensates for the neural networks-based model, allowing a top-down representation of categories characterized by strong specificity and satisfactory results. The only fundamental requirement is that the categories to be represented have corresponding existing frames or synsets, as we will see in the following lines.

A limit already highlighted concerns the inability to distinguish sentences with sensitive content from sentences without. The compositional frames represent the categories of PD; this means that the system identifies the presence or absence of such categories through the presence or absence of the frames that compose them. The result in terms of precision highlights this limitation: we have seen how often categories of PD are identified but are not sensitive because they cannot be attributed to the identifiable subject e.g. Professional.

More specific limits are as follows:

1. No frame or synset for the representation of specific categories. For example, the PDC PIERCING was not modeled, as no matching synsets and frames were found.

2. Some categories reported a high rate of FP, such as PERSONAL POSSESSION (see section 4.3.4). PERSONAL POSSESSION, characterized by the `Possession` frame, is a too-general category that leads to being identified at a semantic level with everything that belongs to the material or abstract subject (*'I have severe pain in my chest'*). Compositional frames seem in general to be more effective in identifying specific categories, as highlighted in the advantages.

Chapter 5

Discussion and Conclusion

Many of these reflections emerged palely during the dissertation. They are discussed here homogeneously (section 5.1), followed by the conclusions which highlight the values of the contributions proposed, the fulfillment of RQs (section 5.2), and a focus on the new horizons that research can open up (section 5.2.2).

5.1 Discussion

In the previous chapters, in particular in section 3.2, the favorite perspective for the exploration of the SID task has been illustrated i.e., the hybrid approach that observed the combination of deep learning with knowledge and sub-symbolical-based methods. Each proposed contribution has been supported by an analysis of the results obtained and a discussion of the advantages and disadvantages of the resource or model in question.

In this section, we want to present a broader discussion, which could highlight the originality and usability of the proposed contributions and which opens the perspective to problems and horizons to be explored and deepened. We will therefore proceed with questions and answers based on scientific evidence or, again, further questions, which can envisage future works.

How do the approaches contributed to the state-of-the-art?

The first evidence that we can underline, concerning the state-of-the-art, is the creation of a labeled resource for the domain of PDCs, freely available upon the signature of an ethical use agreement. So far, in the literature, no resources with this type of label and easily downloadable and usable have been highlighted. The labels are sentence-level and - compared to the state-of-the-art resources e.g., Wikipedia dataset, Pastebin, and social media post dataset (see the corpora report, in Appendix B.1) - cover a large number of PDCs.

Looking at the overview of SID works (Appendix A.1), let's examine those dealing with the domain of personal information, working on the English language and dating back to recent years, from 2019 onwards. By applying these criteria, the studies involved are considerably reduced and in any case, present substantial differences concerning the contribution presented here.

Battaglia et al. [12] work at a post level and aim to discriminate between sensitive and non-sensitive content. The NLP approach to word-embeddings reaches 68% of F1 score. Guo et al. [64] work instead at a sentence-level by implementing a BERT-like context-aware model. It is important to note, however, that the categories analyzed are PII, such as name, address, date of birth, social security number (SSN), and telephone number. The F1 score reaches the

99.5%. Despite the work of Korba et al. [87] dates back to 2008, it must be mentioned for the hybrid approach based on NLP techniques, but at the same time on the rules of relationship between the identified entities. The categories investigated are, as above, PII and - working on a word-level - it reaches a 93.3% of accuracy.

The SID model contributes by suggesting a completely new approach to the problem and working on a number of categories that are firstly larger than the categories in the literature, and conceptually more complex than the PII investigated in previous studies.

How can broad boundary categories be handled?

Generic PDCs have been included in the PRIVAFRAME modeling, identified as broad-boundaries categories. They can be found in Appendix D.1. These categories are not present in SPeDaC 3 and have not been tested. However, it is possible through some examples to observe their extended conceptual nature:

1. BEHAVIORAL: '*Information about Behavior or activity*';

2. INTENTION: '*Information about intentions*';

3. INTEREST: '*Information about interests*'.

These 3 PDCs have been modeled in PRIVAFRAME as follows:

```
1)
dpv: Behavioral owl: equivalentClass fscore: Conduct
2)
dpv: Intention owl: equivalentClass fscore: Purpose
3)
dpv: Interest owl: equivalentClass [owl: intersectionOf (fscore:
EmotionByStimulus fscore: EmotionDirected)].
```

The compositional frames to describe these categories can result not being very effective in automatic identification. Two possible ways to address the problem are hypothesized:

- **Deconstruction of broad-boundaries PDCs.** Let's take as an example the BEHAVIORAL category, a subclass of the EXTERNAL macro-category. First of all, let's consider when a *behavior* or an *activity* can be declined in different contexts e.g., online behavior, social behavior, physical behavior, etc. The DPV foresees some of these specifications; the subclasses of PDC BEHAVIORAL are presented in Fig. 5.1.

  Most of these sub-categories fall within those not identifiable through textual elements, especially those related to online behavior e.g., LINK CLICKED or BROWSING BEHAVIOR. The only specific PDC analyzed is PERFORMANCE AT WORK, whose context is identifiable and is part of the individual's professional information. Among the BEHAVIORAL subclasses, however, the following have been considered broad-boundaries: ATTITUDE, PERSONALITY, and DEMEANOR, as they generally refer to the individual behaviors of an identifiable subject, and often cannot be deduced directly and with certainty from the text. We can derive the personality of an individual ('*Information about personality*'), based on linguistic and non-linguistic behaviors that enact and the subsequent

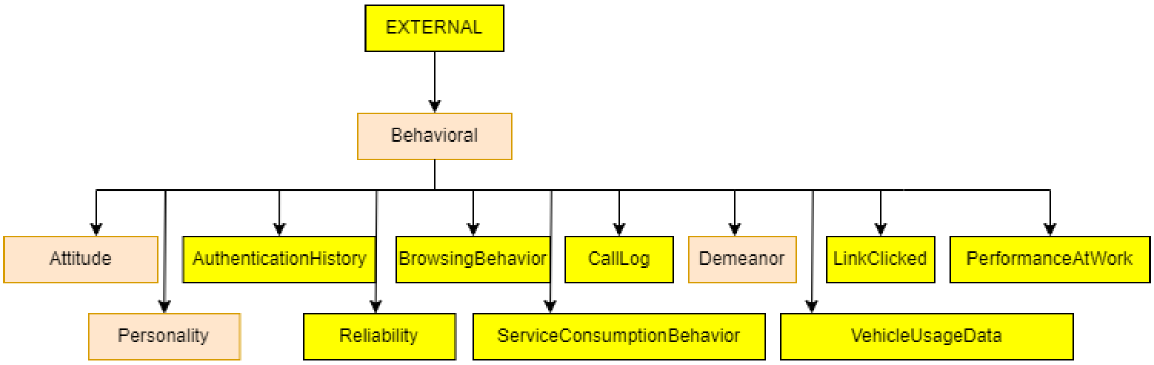


Figure 5.1: Subclasses of :EXTERNAL:BEHAVIORAL PDCs

categorization in terms of the Big Five personality traits, but it is difficult to uniquely retrieve from text information on personality traits.

These concepts could be investigated through our model only by finding a conceptual specification first. Just as a subclass of BEHAVIORAL can mean the individual's performance at a professional level, in the same way, the concept of ATTITUDE could be treated contextually e.g., attitude at work, attitude at school, etc.

To resume, these PDCs are extremely complex concepts that would undoubtedly require an in-depth detailed conceptual study.

- **Expansion of LUs.** The PRIVAFRAME modeling is based on frames present in existing resources and on their composition by logical relationships. Frames aim to describe semantic situations that are often much more generic than the specificity required by the PDCs. If this specificity is not satisfied in the creation of new frames, consequently to the composition between them, as demonstrated, a complementary way could consist in the reuse of the existing frames by extending them; in other words, expanding the LUs that characterize them and creating new PDCs frames from scratch.

  Starting from broad-boundary categories conceptually characterized by a strong vagueness, the expansion of frames such as `Conduct`, associated with the PDC BEHAVIORAL, adding LUs that can help in the pragmatic identification of this concept on text, can be an interesting exploration.

  The creation of frames from the reuse of existing frames and the extension of LUs can also be considered a solution to the PDCs taken into consideration for PRIVAFRAME which, however, have not found an adequate representation through the compositional approach, such as PIERCING or TATTOO.

Would it be possible to extend PRIVAFRAME *to the identification of sensitive and non-sensitive sentences?*

As seen, PRIVAFRAME is able to identify the specific sensitive categories, while the discrimination between sensitive and non-sensitive sentences is entrusted to neural networks. The deep learning approach can help in particular to manage specific ambiguous cases e.g., ironic contexts, expressions of hopes and desires that make the content non-sensible. However, an ontological extension of PRIVAFRAME could be imagined to make the model not only able to identify the PDCs, but also to elaborate the relationships to the other elements of the sentence. Taking up the definition of SPD (see section 2.2), we know that the relationship between the potential sensitive data and the identifiable subject is an essential criterion. The expansion could therefore consist in the introduction of thematic roles of frames e.g., the indication of whom the sensitive data is referred to. FRED, in addition to automatically extract the frames, is able to give information on the thematic roles of the sentences. A similar approach has been already implemented in Sentilo [49, 140]. Sentilo is a tool for automatic sentiment recognition. We can start from the *opinion* definition: '*an opinion can be defined as an intentional statement by somebody (holder) on some fact (topic) that is expressed*'. Sentilo gives a formal representation in the RDF graph identifying (i) the holder(s) and (ii) the topic(s) of an opinion sentence. Holder detection, which achieves a 95% F1 score, relies on selected opinion trigger verbs, which imply the presence of a holder. Leveraging FRED's automatic role identification, it is also possible to identify the relations between an opinion and a given event situation directly or indirectly expressed, and the factual impact of an event on a specific role (positive or negative).

Our problem concerns the disambiguation of the subject linked to the PDC. It would not be enough to identify what the subject is, but whether it actually corresponds to an identifiable subject. Assuming that in most cases the identifiable person is the writer, we should verify that in the sentence the PDC is associated with a first singular or plural person, implicitly or explicitly expressed. We could do this for example introducing linguistic triggers associated with the thematic roles of disambiguous agents and expressed in the form of first-person personal pronouns or possessive adjectives (e.g.,'*I have a severe headache*', '*My sister is called Laura*') and, as for Sentilo, creating *ad hoc* graph relations.

Making PRIVAFRAME able to deal with hypothetical, desire, ironic sentences, etc. i.e., intentions that make the potentially identified sensitive content not founded on reality and therefore not to be protected, can prove to be even more complex. However, it is possible to imagine additional ontological layers for this disambiguation; related works are already present in the literature, such as the aforementioned Sentilo, part of whose architecture could be reused for the identification of desired sentences, or rule-based models proposed for the identification of irony and sarcasm [114].

Could it be possible to identify not only the PDCs but also their degree of sensitivity?

Geng et al. [58] proposed a model to measure the sensitive degree to (i) the type of entity, and (ii) the perspective of affirmation (objective-subjective) (see section 2.3).

If the first distinction is the aim of the fine-grained identification model, we find the second one interesting given a further expansion of PRIVAFRAME. Let's go back to the examples, '*Marco has been recognized as having a disability*' (objective sensitive entity, *a*), and '*Marco often suffers from migraines*' (subjective sensitive entity, *b*). An objective statement can be measured, while a subjective one depends on social and cultural factors. First of all, it must be emphasized that this distinction can be only applied to certain categories i.e., the categories that the aforementioned authors define as '*sensitive entities*'; while it is not possible to apply it to the so-called '*quasi-identifying entities*' i.e., to those that do not provide for '*subjective*' structure. There is no subjective way to indicate your home address, age, or weight. If so, the sentence would be hypothetical, dubious, and therefore not identifiable as sensitive. Otherwise, categories such as health (see example above), religion or school can make use of this distinction. '*I received the sacraments of the Catholic Church*' can be measured with a higher score of sensitivity, compared to '*I recently approached Catholicism*'. '*I got a bachelor's degree in literature and a master's in journalism*' can be considered more sensitive than '*I studied for a long time before starting work*'.

Indeed, objective entities may further present different degrees of sensitivity to the specificity of the information attributed to them. In these words, '*I got a bachelor's degree in literature and a master's in journalism*' is more sensitive than '*I got a bachelor's degree and a master*'. Just like Geng et al. [58] did for the restricted domain of the medical field (5 diseases tested), to make PRIVAFRAME sensitive to the degree of sensitivity, we envisage a rule-based approach; for each PDCs considered '*sensitive*', it is necessary to identify recurring linguistic patterns representing both cases and create sets of rules.

Could the model be extended to other languages?

An interesting development of the work is certainly the extension to other languages, in addition to the English language, of the proposed two-fold approach model. In particular, our interest would fall primarily in the Italian language.

Adaptation to the transform-based model would involve the need to find representative corpora: a multilingual extension of SPeDaC. Given the availability of public web corpora in different languages (as we have seen, the TenTen family of corpora covers more than 40 languages), this would not be difficult to realize. At the same time, it is certainly possible to implement a LM for the Italian language.

Similarly, the extension of the top-down model, PRIVAFRAME, would be possible and would not require a training dataset. The structure and the PDCs could be the same even if a literal adaptation is to be excluded. The constructions that occur in representative sentences of PDCs can vary from language to language. The other obstacle to note is that Framester and FrameNet currently make the resource publicly usable in English. Therefore, if we are interested in an extension in Italian, we must first obtain access to the equivalent resource. IFrameNet [11] is a recently developed resource and may not cover the entire domain currently covered by English FrameNet. Some frames of our compositions may therefore not yet have been defined.

5.2 Conclusion

Working on the SID task to ensure the protection of privacy assumes a theoretical complexity from a non-trivial starting point. What do we mean by privacy? What for sensitive data? What are the sensitive data to protect and why? How can these be treated? The approach to the problem, therefore, started by outlining a very precise domain of investigation that could be (i) broad and as adaptable as possible; (ii) of such complexity as to make exploration - at the state-of-the-art - challenging and interesting. We have decided not to deal with basic PD (e.g., credit card numbers, SSN, addresses), which was not only treated with great success in the literature [64] but also have identification tools marketed, e.g., Microsoft [9].

The biggest limitations encountered during the work have been the following:

- The lack of shared resources noted for our domain of interest (in general - in the field of sensitive data - the lack of shared resources is a widespread problem);
- The consequent difficulty in comparing performances with the state-of-the-art.

The contributions aimed in general at:

- Offering datasets that do not directly involve identifiable subjects and are therefore free from direct ethical implications, available for training on other models or for being used as benchmarks;

- On these datasets, measuring the performance of models widely used in SID task literature, based on transformer networks, and finally highlighting their advantages and limitations;
- Proposing an original approach based on a knowledge graph of compositional frames could also contribute as part of a hybrid model. This approach, evaluated on the identification of fine-grained PDCs, is introduced as new to the state-of-the-art. Its main attractions lie in the demand for zero training data and few computational resources. Furthermore, the model is tailor-made and new personal categories can be introduced and modeled with a top-down approach.

We now outline the research questions fulfillment and propose future work to improve, broaden or question these contributions and the generic literature on the SID task.

5.2.1 Fulfillment of ROs

RQ1. We have looked for a theoretical definition of what is meant by sensitive data in sections 2.1.1 and 2.2, reflecting on the complexity that emerges on the theoretical level around the generic concept of privacy, reporting the definitions of a legal nature, finally modeling a definition of sensitive data and in particular of personal sensitive data that could provide useful elements for our automatic identification task. Starting from these definitions we have created a well-defined domain of investigation based on the conceptual and ontological organization work proposed by Pandit [129]. The reference resource, the DPV, is the only and most complete taxonomic resource of PD currently present in the literature and has been extensively described in section 2.4.2. The analysis that leads to the definition of the domain has instead been addressed in section 4.1.2 and leads to the identification of a rather large number of complex PDCs that can be explicitly or implicitly inferred at a textual level.

RQ2. This second research question led us to think of an original approach to the problem compared to the state-of-the-art; at the same time, we were interested in investigating, comparing and combining an approach based on deep learning (an approach on which even the most recent SID are based) -in particular, transformer-based - with logical, top-down approaches that do not require large amounts of training data (chapter 3 and chapter 4 are mainly dedicated to the description of the methodology and the realization of the model).

RQ3. This is the question for which we encountered the most limitations and difficulties (discussing related works - section 2.3 - and available labeled corpora - section 2.5). Furthermore, as it emerges from the evaluation of our models, the absence of a shared benchmark or annotated corpora made difficult a direct comparison with the state-of-the-art in terms of performance measures. We wanted to address the problem by building a labeled resource; on this resource, we conducted our evaluation experiments. The annotated datasets are released and available. We hope they can be used as training or benchmark sets for future approaches to the task.

5.2.2 Future Work

Future work may follow different directions and objectives. Of course, the proposed approaches can be adopted to explore sensitive data domains other than the one investigated here on personal information. It is then possible to identify more defined interventions, which aim to improve the results obtained so far or to expand the functionalities of the model, as well as more wide-ranging works to compare and discuss the approaches adopted here.

Future work to improve or refine the performances of the proposed resource and models are the following:

- The SPeDaC datasets could be quantitatively expanded by collecting more sentences and/or creating sub-corpora that take into account different conversation domains (tweets, Facebook conversations in public pages, blogs and personal web pages, etc.);

- Regarding PRIVAFRAME, PDCs that received a critical identification can be in-depth analyzed and further tested to improve the KG performance. A lexical expansion of some frames could help to create compositional frames apt to represent very specific categories e.g., PIERCING.

Some questions from the discussion in section 5.1 opened the horizons to possible future work that would contribute to the expansion of the resource and whose development hypotheses have already been discussed above:

- The multilingual expansion of the models;

- The transformer-based classifier is currently able to reason on sentence-level spans. To this classification, a finer, token-level classification could be added. This would result in a new dataset annotation. An encoding format following the BIO schema could be adopted.

- SPeDaC could be tested on other state-of-the-art models. Specifically, we are considering the fine-tuning of a sentence-transformer model [141] e.g., LaBSE (Language-Agnostic BERT Sentence Embedding) [41]. Sentence-transformer models work by giving the input text to a pre-trained transformer model, able to extract contextualized word embeddings. A pooling layer averages the word embeddings extracted to get a fixed length and high dimensional vector. Sentence transformers seem to drastically reduce the time computation if compared to BERT-like models.

In particular, they concern with the PRIVAFRAME expansion:

- An in-depth study, analysis and modeling of broad-boundaries PDCs in PRIVAFRAME. As presented, these categories have currently received the first modeling, but have not been tested and, as discussed, would need a more in-depth study at a top-down level;

- Introduction of ontological reasoning in PRIVAFRAME, to test its ability not only to work on identifying fine-grained PDCs but also to discriminate sensitive *vs* non-sensitive content. A follow-up to this work concerns the identification through inference rules of expressions able to cancel the potential sensitivity of the statements (hypothetical, ironic, sarcastic, etc.);

- Creation of an additional layer to the PRIVAFRAME fine-grained identification, to confer a measure of sensitivity to the identified PDC. This involves analyzing each interested category, identifying recurring patterns, and creating rules.

Future work to contribute to the SID task:

- We have seen how PRIVAFRAME, on some PDCs in particular, reported a high number of FP. For automatic frame identification, it would be interesting to compare another hybrid approach recently proposed in the literature: PAFIBERT [166]. It is a solution that intends to combine the LMs representation capability of BERT with a position-based attention mechanism to capture target-specific contextual information: the model identifies different instances of the target based on their positions in the sentence, in other words, it's able to retrieve the frames in relation to the context. PAFIBERT proposes a frame filtering mechanism based on:

  - Frame filtering by LUs: gold selected LUs are used to retrieve candidate frames in sentences. The training data present examples sentences of the LUs;
  - Frame filtering by targets: when the golden LUs are not marked, a pre-processing of the sentence helps in identifying the potential candidate frames for frame filtering.

  Indeed, we have tried to concretely implement this idea, but so far we have not been able to get the model presented. It would be interesting to verify if the contextual awareness in the automatic frame identification and the filtering LUs system could help the model to improve in precision.

Finally, the twofold model can be already made usable by developing a Web interface.

Bibliography

[1] Abdiansah, A., Wardoyo, R.: *Time complexity analysis of support vector machines (svm) in libsvm*. International Journal of Computer Applications, 128, 3, pp. 28–34 (2015) https://doi.org/10.5120/ijca2015906480

[2] Adhikari, K., Panda, R.: *Users' Information Privacy Concerns and Privacy Protection Behaviors in Social Networks*. Journal of Global Marketing, 1, pp. 1-15 (2018) https://doi.org/10.1080/08911762.2017.1412552

[3] Agrawal, R., Imieliński, T., Swami, A.: *Mining association rules between sets of items in large databases*. SIGMOD Rec. 22, 2, pp. 207–216 (1993) https://doi.org/10.1145/170036.170072

[4] Allahyari, M., Pouriyeh, S., Assefi, M., et al.: *A brief survey of text mining: classification, clustering and extraction techniques*. arXiv (2017) https://arxiv.org/abs/1707.02919

[5] Andrade, N.N.G.: *Data Protection, Privacy and Identity: Distinguishing Concepts and Articulating Rights*. Privacy and Identity Management for Life - IFIP Advances in Information and Communication Technology, 352, pp. 90-107 (2011) https://doi.org/10.1007/978-3-642-20769-3_8

[6] Androutsopoulou, A., Karacapilidis, N., Loukis, E., et al.: *Transforming the communication between citizens and government through AI-guided chatbots*. Government Information Quarterly, 36, 2, pp. 358-367 (2019) https://doi.org/10.1016/j.giq.2018.10.001

[7] Bahdanau, D., Cho, K., Bengio, Y.: *Neural Machine Translation by Jointly Learning to Align and Translate*. 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, Conference Track Proceedings (2015) http://arxiv.org/abs/1409.0473

[8] Baker, C.F., Fillmore, C.J., Lowe, J.B.: *The Berkeley FrameNet Project*. COLING-ACL '98 Proceedings of the Conference, held at the University of Montreal, Association for Computational Linguistics, pp. 86-90 (1998) https://doi.org/10.3115/980845.980860

[9] Balzer, A., Mowatt, D., Woulfe, M.: *Protecting personally identifiable information (pii) using tagging and persistence of pii*. US Patent (2021) https://patents.justia.com/patent/10885225

[10] Bartlett, F. C.: *Remembering: A study in experimental and social psychology*. Cambridge University Press. (1967)

[11] Basili, R., Brambilla, S., Croce, D., et al.: *Developing a large scale framenet for italian: the iframenet experience*. CLiC-it, Rome, pp. 59-64 (2017) http://ceur-ws.org/Vol-2006/paper079.pdf

[12] Battaglia, E., Bioglio, L., Pensa, R. G.: *Towards Content Sensitivity Analysis.* Berthold, M., Feelders, A., Krempl, G. (eds) Advances in Intelligent Data Analysis XVIII, Lecture Notes in Computer Science, 12080, Springer, Cham (2020) https://doi.org/10.1007/978-3-030-44584-3_6

[13] Bench-Capon, T. J. M., Visser, P. R. S.: *Open texture and ontologies in legal information systems.* Database and Expert Systems Applications, 8th International Conference, DEXA '97, Proceedings, pp. 192-197 (1997)

[14] Berardi, G., Esuli, A., Macdonald, C. et al.: *Semi-automated text classification for sensitivity identification.* Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, Association for Computing Machinery (ACM), pp. 1711–1714 (2015) https://doi.org/10.1145/2806416.2806597

[15] BeVier, L. R.: *Information About Individuals in the Hands of Government: Some Reflections on Mechanisms for Privacy Protection.* 4 Wm. Mary Bill Rts. J. 455 (1995)

[16] Bilal, M., Almazroi, A. A.: *Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews.* Electronic Commerce Research (2022) https://doi.org/10.1007/s10660-022-09560-w

[17] Bisong, E.: *Logistic Regression.* Building Machine Learning and Deep Learning Models on Google Cloud Platform, Apress, Berkeley, CA (2019) https://doi.org/10.1007/978-1-4842-4470-8_20

[18] Blei, D., Ng, A., Jordan, M.: *Latent dirichlet allocation.* The Journal of machine Learning research, 3, pp. 993–1022 (2003)

[19] Bokaie, H.M., Irwin, R., Serge, E.: *Identifying and classifying thirdparty entities in natural language privacy policies.* Proceedings of 2nd Workshop Privacy, pp. 18–27 (2020) https://doi.org/10.18653/v1/2020.privatenlp-1.3

[20] Breuker, J., Valente, A., Winkels, R. *Legal Ontologies in Knowledge Engineering and Information Management.* Artificial Intelligence and Law, 12, pp. 241-277 (2004) https://doi.org/10.1007/s10506-006-0002-1

[21] Briskilal, J., Subalalitha, C. N.: *An ensemble model for classifying idioms and literal texts using BERT and RoBERTa.* Information Processing Management, 59 (2022) https://doi.org/10.1016/j.ipm.2021.102756

[22] Cavoukian, A.: *7 Foundational Principles.* Office of the Information and Privacy Commissioner (2009)

[23] Celli, F., Pianesi, F., Stillwell, D., et al.: *Workshop on computational personality recognition: shared task.* Proceedings of the International AAAI Conference on Web and Social Media, 7, 2, pp. 2-5 (2013)

[24] Chang, C. C., Lin, C. J.: *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2, 3 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last access January 18, 2023)

[25] Cho, K., van Merrienboer, B., Gulcehre, C., et al.: *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734 (2014) https://doi.org/10.3115/v1/D14-1179

[26] Chow, R., Philippe, G., Staddon, J.: *Detecting privacy leaks using corpus-based association rules.* Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, Association for Computing Machinery, New York United States, pp. 893-901 (2008) https://doi.org/10.1145/1401890.1401997

[27] Church, K., Gale, W., Hanks, P., et al.: *Using Statistics in Lexical Analysis.* Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon, pp. 115-164 (1991)

[28] Coavoux, M., Narayan, S., Cohen, S. B.: *Privacy-preserving neural representations of text.* Proceedings of Conference Empirical Methods Natural Language Processing, pp. 1–10 (2018) https://doi.org/10.48550/arXiv.1808.09408

[29] Coppola, B., Gangemi, A., Gliozzo, A., et al.: *Frame Detection over the Semantic Web.* ESWC Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 5554, pp. 126–142 (2009) https://doi.org/10.1007/978-3-642-02121-3_13

[30] Cortes, C., Vapnik, V.: *Support-vector networks.* Machine Learning, pp. 273–297 (1995) https://doi.org/10.1007/BF00994018

[31] Cumby, C., Ghani, R.: *Inference Control to Protect Sensitive Information in Text Documents.* ACM SIGKDD Workshop on Intelligence and Security Informatics, 5, pp. 1-7 (2010) https://doi.org/10.1145/1938606.1938611

[32] De Campos, L.M., Moral, S.: *Learning rules for a fuzzy inference model.* Fuzzy Sets and Systems, 59, 3, pp. 247-257 (1993) https://doi.org/10.1016/0165-0114(93)90470-3

[33] De Hert, P. J. A., Gutwirth, S.: *Privacy, data protection and law enforcement: Opacity of the individual and transparency of power.* Claes, E., Duff, A., Gutwirth, S.: Privacy and the criminal law, Intersentia, pp. 61-104 (2006)

[34] Devlin, J., Chang, M.W., Lee, K., et al.: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, 1 , pp. 4171–4186 (2019)

[35] Dias, M., Bon, J., Ferreira, J.C., et al.: *Named Entity Recognition for Sensitive Data Discovery in Portuguese.* Applied Sciences, 10, 7, p. 2303 (2020) https://doi.org/10.3390/app1007230

[36] Dıaz, N., Cuellar, M.P., Lilius, J., et al.: *A fuzzy ontology for semantic modelling and recognition of human behaviour*, 66, pp. 46-60 (2014) https://doi.org/10.1016/j.knosys.2014.04.016

[37] Fazzinga, B., Galassi, A., Torroni, P.: *A Preliminary Evaluation of a Privacy-Preserving Dialogue System.* Fifth Workshop on Natural Language for Artificial Intelligence (2021)

[38] Fellbaum, C.: *A Semantic Network of English: The Mother of All WordNets.* Computers and the Humanities, 32, pp. 209–220 (1998) https://doi.org/10.1007/978-94-017-1491-4_6

[39] Fellbaum, C.D.: *WordNet : an electronic lexical database.* Language, 76, p. 706 (2000)

[40] Fellbaum, C.D.: *Harmonizing WordNet and FrameNet.* Loftsson, H., Rögnvaldsson, E., Helgadóttir, S.: Advances in Natural Language Processing. NLP 2010, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 6233 (2010) https://doi.org/10.1007/978-3-642-14770-8_2

[41] Feng, F. et al.: *Language-agnostic bert sentence embedding.* arXiv (2020) https://doi.org/10.48550/arXiv.2007.01852

[42] Ferrucci, D., Lally. A.: *Building an example application with the Unstructured Information Management Architecture.* IBM Systems Journal, 43, 3, pp. 455-475 (2004) https://doi.org/10.1147/sj.433.0455

[43] Fillmore, C. J.: *Frame semantics and the nature of language.* Annals of the New York Academy of Sciences, 280, pp. 20-32 (1976) https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

[44] Fillmore C.J., Baker C.: *Frame semantics for Text Understanding.* Proceedings of NAACL 2001, WordNet and Other Lexical Resources Workshop, Pittsburgh, Pennsylvania (2001)

[45] Freudenthal, D., Pine, J., Gobet, F.: *On the Utility of Conjoint and Compositional Frames and Utterance Boundaries as Predictors of Word Categories.* Proceedings of the Annual Meeting of the Cognitive Science Society, 30, pp. 1947-1952 (2008)

[46] Freund, Y., Schapire, R. E.: *Experiments with a new boosting algorithm.* ICML, 96, pp. 148–156 (1996)

[47] Gagliardi, G.: *Inter-Annotator Agreement in linguistica: una rassegna critica.* Computational Linguistics CLiC-it, 206 (2018)

[48] Gangemi, A., Navigli, R., Velardi, P.: *The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet.* Proceedings of International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003). Catania, Sicily (Italy), pp. 820–838 (2003) https://doi.org/10.1007/978-3-540-39964-3_52

[49] Gangemi A, Presutti V, Reforgiato Recupero D.: *Frame-based detection of opinion holders and topics: a model and a tool.* IEEE Computing Intelligence, 9, 1, pp. 20–30 (2014) https://doi.org/10.1109/MCI.2013.2291688

[50] Gangemi A., Alam M., Asprino L. et al.: *A Wide Coverage Linguistic Linked Data Hub.* EKAW, Springer International Publishing, pp. 239–254 (2016) https://doi.org/10.1007/978-3-319-49004-5_16

[51] Gangemi, A., Presutti, V., Reforgiato Recupero, D., et al.: *Semantic Web Machine Reading with FRED.* Semantic Web, IOS Press, pp. 2210-4968 (2016) https://doi.org/10.3233/SW-160240

[52] Gangemi, A., Mehwish, A., Presutti, V.: *Amnestic Forgery: An Ontology of Conceptual Metaphors*. Borgo, S., Hitzler, P., Kutz, O.: Formal Ontology in Information Systems, Proceedings of the 10th International Conference, FOIS 2018, Cape Town, South Africa, 19-21 September 2018, Frontiers in Artificial Intelligence and Applications, 306, pp. 159-172, IOS Press (2018) https://doi.org/10.3233/978-1-61499-910-2-159

[53] Gangemi, A.: *Closing the Loop Between Knowledge Patterns in Cognition and the Semantic Web*. Semantic Web, IOS Press, 1 Jan., pp. 139–151 (2020) https://doi.org/10.3233/SW-190383

[54] Garcia, A. X.: *Identifying Sensitive Information in Text Using an Ontological Knowledge Base Information Extraction and Logical Inferencing*. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2017)

[55] Garcia Pablos, A., Perez, N., Cuadros, M.: *Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT*. Proceedings The 12th Language Resources and Evaluation Conference, pp. 4486–4494 (2020)

[56] Garijo, D., Gil, Y.: *Augmenting PROV with plans in P-Plan: scientific processes as linked data*. CEUR Workshop Proceedings (2012)

[57] Genetu A., Tegegne T.: *Designing Sensitive Personal Information Detection and Classification Model for Amharic Text*. International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp. 54-58 (2021) https://doi.org/10.1109/ICT4DA53266.2021.9672227

[58] Geng, L., You Y., Liu, H. et al.: *Privacy measures for free text documents: Bridging the gap between theory and practice*. Proceedings of the 8th International Conference on Trust, Privacy and Security in Digital Business,TrustBus '11, Springer-Verlag, Berlin Heidelberg, pp. 161–173 (2011) https://doi.org/10.1007/978-3-642-22890-2_14

[59] Gildea, D., Jurafsky, D.: *Automatic labeling of semantic roles*. Comput. Linguist., 28, 3, pp. 245–288 (2002) https://doi.org/10.1162/089120102760275983

[60] Graves, A., Schmidhuber, J.: *Framewise phoneme classification with bidirectional LSTM and other neural network architectures*. Neural Networks, 18, 5, pp. 602-610 (2005) https://doi.org/10.1016/j.neunet.2005.06.042

[61] Grice, H. P.: *Utterer's Meaning and Intentions*. Philosophical Review, 78, 147-77 (1969)

[62] Gruber, T. R.: *A translation approach to portable ontology specifications*. Knowledge Acquisition, 5, 2, pp. 199-220 (1993) https://doi.org/10.1006/knac.1993.1008

[63] Guarino, N.: *Ontologies and knowledge bases: towards a terminological clarification*. Towards Very Large Knowledge BasesPublisher, Amsterdam, IOS Press Editors, Mars, N.J.I, pp. 25-32 (1995)

[64] Guo, Y., Liu, J., Tang, W. et al.: *Exsense: Extract sensitive information from unstructured data*. Computers & Security, 102, pp. 102-156 (2021) https://doi.org/10.1016/j.cose.2020.102156

[65] Hage, J., *A Theory of Legal Reasoning and A Logic To Match*. Artificial Intelligence and Law, 4, pp. 3-4 (1996) https://doi.org/10.1007/BF00118493

[66] Harris, Z.: *Distributional structure*. Word, 10, pp. 146-162 (1954) https://doi.org/10.1007/978-94-009-8467-7_1

[67] Hart, H. L. A. *Positivism and the Separation of Law and Morals*. Harvard Law Review, 71, 4, pp. 593–629 (1958) https://doi.org/10.2307/1338225

[68] Hart, M., Johnson, R., Manadhata, P.: *Text classification for data loss prevention*. Proceedings of the 11th International Conference on Privacy Enhancing Technologies, PETS '11, Springer Verlag, Berlin Heidelberg, pp. 18-37 (2011) https://doi.org/10.1007/978-3-642-22263-4_2

[69] Hayes, A.F., Krippendorff, K.: *Answering the Call for a Standard Reliability Measure for Coding Data*. Communication Methods and Measures, 1, 1, pp. 77-89 (2007) https://doi.org/10.1080/19312450709336664

[70] Hendrickx, I., van Waterschoot, J., Khan, A., et al.: *Take Back Control: User Privacy and Transparency Concerns in Personalized Conversational Agents*, Joint Proceedings of the ACM IUI 2021 Workshops (2021)

[71] Hochreiter, S., Schmidhuber, J.: *Long Short-term Memory*. Neural computation, 9, pp. 1735-80 (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[72] Hoekstra, R., Groth, P.: *PROV-O-Viz - Understanding the Role of Activities in Provenance*. Ludäscher, B., Plale, B.: Provenance and Annotation of Data and Processes, IPAW 2014, Lecture Notes in Computer Science, 8628, Springer, Cham (2015) https://doi.org/10.1007/978-3-319-16462-5_18

[73] Hogan, A., Blomqvist, E., Cochez, M.: *Knowledge Graphs*. Knowledge Graphs. ACM Computing Survey 54, 4, Article 71 (2021)

[74] Honnibal, M., Montani, I.: *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental* (2017)

[75] Hüllermeier, E.: *Fuzzy sets in machine learning and data mining*. Applied Soft Computing, 11, 2, pp. 1493-1505 (2011) https://doi.org/10.1016/j.asoc.2008.01.004

[76] Hüllermeier, E.: *From knowledge-based to data-driven modeling of fuzzy rule-based systems: A critical reflection*. ArXiv (2017) https://doi.org/10.48550/arXiv.1712.00646

[77] Imocrante, M., Zanetti, L. *WHY OPEN TEXTURE?*. APhEX Portale italiano di filosofia analitica, 21 (2020)

[78] Islam, A. C., Greenstadt, R., Walsh, J.: *Privacy detective: Detecting private information and collective privacy behavior in a large social network*. Proceedings of the 13th Workshop on Privacy in the Electronic Society,WPES '14, Association for Computing Machinery, New York United States, pp. 35-46 (2014) https://doi.org/10.1145/2665943.2665958

[79] Jakubicek, M., Kilgariff, A., Kovar, V., et al.: *The TenTen Corpus Family*. 7th International Corpus Linguistics Conference CL 2013, Lancaster, 2013, pp. 125-127 (2013)

[80] Kao, D., Archer, N.P.: *Abstraction in conceptual model design.* International Journal of Human-Computer Studies, 46, 1, pp. 125-150 (1997) https://doi.org/10.1006/ijhc.1996.0086

[81] Kearns, M.J.: *The computational complexity of machine learning.* MIT press (1990)

[82] Kilgarriff, A., Rychly, P., Smrz, P. et al.: *The Sketch Engine.* Proceedings Euralex, Lorient, France (2004)

[83] Kim, Y.: *Convolutional Neural Networks for Sentence Classification.* arXiv (2014) https://doi.org/10.48550/arXiv.1408.5882

[84] Klie, J.C., Bugert, M., Boullosa, B. et al.: *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.* Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico. Association for Computational Linguistics, pp. 5–9 (2018)

[85] Klimt, B., Yang, Y.: *Introducing the Enron Corpus.* CEAS (2004)

[86] Klimt, B., Yang, Y.: *The Enron Corpus: A New Dataset for Email Classification Research.* Boulicaut, J. F., Esposito, F., Giannotti, F., et al.: Lecture Notes in Computer Science, Machine Learning: ECML, 3201. Springer, Berlin, Heidelberg (2004) https://doi.org/10.1007/978-3-540-30115-8_22

[87] Korba, L., Wang, Y., Geng, L., et al.: *Private Data Discovery for Privacy Compliance in Collaborative Environments.* Proceedings of the Fifth International Conference on Cooperative Design, Visualization and Engineering (CDVE 2008), Palma de Mallorca, Spain, pp. 21-25 (2008) https://doi.org/10.1007/978-3-540-88011-0_18

[88] Krennmayr, T.: *Top-down versus bottom-up approaches to the identification of metaphor in discourse.* metaphorik.de, 24, pp. 7-36 (2013)

[89] Krippendorff, K.: *Reliabilty in content analysis: some common misconceptions and recommemdations.* Human Communication Research, 30, 3, pp. 411-433 (2004) https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[90] Kshirsagar, M., Thomson, S., Schneider, N., et al.: *Frame-Semantic Role Labeling with Heterogeneous Annotations.* Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2, Beijing, China, pp. 218–224 (2015) https://doi.org/10.3115/v1/P15-2036

[91] Kulkarni, P., Cauvery, N. K.: *Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique.* International Journal of Advanced Computer Science and Applications, West Yorkshire, 12, 9 (2021) https://doi.org/10.14569/IJACSA.2021.0120957

[92] Lafferty, J., Mccallum, A., Pereira, F.: *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282-289 (2001)

[93] Landis, J. R., Koch, G. G.: *The Measurement of Observer Agreement for Categorical Data.* Biometrics, 33, 1, pp. 159–174 (1977) https://doi.org/10.2307/2529310

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

[95] LeCun, Y., Boser, B., Denker, J. et al.: *Handwritten Digit Recognition with a Back-Propagation Network.* Advances in Neural Information Processing Systems,Morgan-Kaufmann, 2 (1989)

[96] Lewis, D.: *Naive Bayes at forty: The independence assumption in information retrieval.* Machine Learning: ECML-98, Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, Springer, Berlin, pp. 4-15 (1998) https://doi.org/10.1007/BFb0026666

[97] Li, N., Li, T. Venkatasubramanian, S.: *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.* Proceedings of the 23rd International Conference on Data Engineering, Istanbul, Turkey, pp. 106-115 (2007) https://doi.org/10.1109/ICDE.2007.367856

[98] Li, F., Xie, W., Wang, X. et al.: *Research on Optimization of Knowledge Graph Construction Flow Chart.* IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 1386-1390 (2020) https://doi.org/10.1109/ITAIC49862.2020.9338900

[99] Lin, Z., Feng, M., Nogueira dos Santos, C., et al.: *A Structured Self-attentive Sentence Embedding.* ArXiv (2017) https://doi.org/10.48550/arXiv.1703.03130

[100] Lin Y., Xu G., Xu G. et al.: *Sensitive Information Detection Based on Convolution Neural Network and Bi-Directional LSTM.* IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, pp. 1614-1621 (2020) https://doi.org/10.1109/TrustCom50675.2020.00223

[101] Liu, Y., Ott, M., Goyal, N., et al.: *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* ArXiv (2019) https://doi.org/10.48550/arxiv.1907.11692

[102] Loshchilov, I., Hutter, F.: *Decoupled Weight Decay Regularization.* Proceedings of the Seventh International Conference on Learning Representations (2019)

[103] Lukács, A.: *What is Privacy? The History and Definition of Privacy.* (2016)

[104] Macdonald, C., McDonald, G., Ounis, I.: *Enhancing sensitivity classification with semantic features using word embeddings.* European Conference on Information Retrieval, Springer International Publishing, pp. 450-463 (2017) https://doi.org/10.1007/978-3-319-56608-5_35

[105] Machanavajjhala A., Gehrke, J., Kifer. D.: *l-Diversity: Privacy Beyond k-Anonymity.* Proceedings of the 22nd International conference on Data Engineering, Atalanta, USA, pp. 24-24 (2006) https://doi.org/10.1109/ICDE.2006.1

[106] Majeed, A., Lee, S.: *Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey.* IEEE Access, pp. 1-35 (2020) https://doi.org/10.1109/ACCESS.2020.3045700

[107] Manning, C. D., Surdeanu, M., Bauer, J. et al.: *The Stanford CoreNLP natural language processing toolkit*. Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60 (2014) https://doi.org/10.3115/v1/P14-5010

[108] McDonald, G., Macdonald, C., Ounis, I. et al.: *Towards a classifier for digital sensitivity review*. Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval, ECIR 2014, Springer International Publishing, , pp. 500–506 (2014) https://doi.org/10.1007/978-3-319-06028-6_48

[109] Mehwish, A., Gangemi, A., Presutti, V. et al.: *Semantic role labeling for knowledge graph extraction from text*. Progress in Artificial Intelligence, 10, pp. 309–320 (2021) https://doi.org/10.1007/s13748-021-00241-7

[110] Mikolov, T., Sutskever, I., Chen, K. et al.: *Distributed Representations of Words and Phrases and their Compositionality*. Advances in Neural Information Processing Systems, 26, pp. 3111–3119 (2013)

[111] Miller, A. R.: *The assault on privacy : computers, data banks, and dossiers*. Ann Arbor: University of Michigan Press (1971)

[112] Miller, G.A.: *WordNet: A Lexical Database for English*. Communications of the ACM, 38, 11, pp. 39-41 (1995) https://doi.org/10.1145/219717.219748

[113] Minsky, M.: *A Framework for Representing Knowledge*. Winston, P.: The Psychology of Computer Vision, New York, McGraw Hill, pp. 211-277 (1975) https://doi.org/10.1515/9783110858778-003

[114] Mladenovic, M., Krstev, C., Mitrović, J. et al.: *Using Lexical Resources for Irony and Sarcasm Classification*. Proceedings of the 8th Balkan Conference in Informatics, 13, pp. 1-8 https://doi.org/10.1145/3136273.3136298

[115] Müller, R., Kornblith, S., Hinton, G.: *When does label smoothing help?*. 422, Red Hook, NY, USA, Curran Associates Inc. (2019)

[116] Mulligan, D. K., Koopman, C., Doty, N.: *Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy*. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 374, pp. 2083 (2016) https://doi.org/10.1098/rsta.2016.0118

[117] Murthy, S., Bakar, A., Rahim, F. et al.: *A Comparative Study of Data Anonymization Techniques*. IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), pp. 306-309 (2019) https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00063

[118] Navigli, R., Ponzetto, S.P.: *BabelNet: Building a Very Large Multilingual Semantic Network*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, pp. 216–225 (2010)

[119] Neerbek, J., Assent, I., Dolog, P.: *Detecting complex sensitive information via phrase structure in recursive neural networks*. Advances in Knowledge Discovery and Data Mining, PAKDD '18, pp. 373-385. Springer, Cham (2018) https://doi.org/10.1007/978-3-319-93040-4_30

[120] Neerbeck, J.,: *Sensitive information detection: Recursive neural networks for encoding context*. PhD Thesis. Aarhus University Denmark (2020)

[121] Neerbek, J., Eskildsen, M., Dolog, P., Assent, I.: *A real-world data resource of complex sensitive sentences based on documents from the monsanto trial*. Proceedings of The 12th Language Resources and Evaluation Conference, pp. 1258–1267 (2020)

[122] Noever, D.: *The Enron Corpus: Where the Email Bodies are Buried?*, arXiv (2020) https://doi.org/10.48550/arxiv.2001.10374

[123] Nuzzolese, A. G., Gangemi, A., Presutti, V. et al.: *Automatic Typing of DBpedia Entities*. Proceedings of the International Semantic Web Conference (ISWC), Boston, MA, US (2012) https://doi.org/10.1007/978-3-642-35176-1_5

[124] Obeid, J. S., Heider, P. M., Weeda, E. R., et al.: *Impact of De-Identification on Clinical Text Classification Using Traditional and Deep Learning Classifiers*. Stud Health Technol Inform., 264, pp. 283-287 (2019) https://doi.org/10.3233/SHTI190228

[125] Oltramari, A., Piraviperumal, D., Schaub, F. et al.: *PrivOnto: A semantic framework for the analysis of privacy policies*. Semantic Web, 9, pp. 1-19. (2017) https://doi.org/10.3233/SW-170283

[126] Otmakhova, Y., Shin, H.: *Do We Really Need Lexical Information? Towards a Top-down Approach to Sentiment Analysis of Product Reviews*. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado. Association for Computational Linguistics, pp. 1559–1568 (2015) https://doi.org/10.3115/v1/N15-1179

[127] Palmirani, M., Martoni, M., Rossi, A. et al.: *PrOnto: Privacy Ontology for Legal Reasoning*. Kő, A., Francesconi, E.: Lecture Notes in Computer Science, (eds) Electronic Government and the Information Systems Perspective, EGOVIS 2018, 11032, Springer, Cham (2018) https://doi.org/10.1007/978-3-319-98349-3_11

[128] Pandit, H. J., Fatema, K., O'Sullivan, D., et al.: *GDPRtEXT - GDPR as a Linked Data Resource*. The Semantic Web. ESWC 2018. Lecture Notes in Computer Science, 10843 (2018) https://doi.org/10.1007/978-3-319-93417-4_31

[129] Pandit H. J., Polleres A., Bos B. et al.: *Creating a vocabulary for data privacy: the first-year report of data privacy vocabularies and controls community group (DPVCG)*. OTM 2019 Conferences Proceedings. Springer International Publishing, pp. 714-730 (2019) https://doi.org/10.1007/978-3-030-33246-4_44

[130] Pandit, H. J., Debruyne, C., O'Sullivan, et al.: *GConsent - A Consent Ontology Based on the GDPR*. The Semantic Web, ESWC 2019, Lecture Notes in Computer Science, 11503 (2019) https://doi.org/10.1007/978-3-030-21348-0_18

[131] Pandit H. J.: *Representing Activities associated with Processing of Personal Data and Consent using Semantic Web for GDPR Compliance*. Trinity College Dublin, School of Computer Science & Statistics, Ph.D. Thesis (2020)

[132] Pedregosa, F., Varoquaux, G., Gramfort, A. et al.: *Scikit-learn: machine learning in Python*. Machine Learning for Evolution Strategies, 12, pp. 2825–2830 (2011) https://doi.org/10.48550/arXiv.1201.0490

[133] Pennington, J., Socher, R., Manning, C.: *GloVe: Global Vectors for Word Representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 1532-1543 (2014)

[134] Peters, M. E., Neumann, M., Iyyer, M. et al.: *Deep contextualized word representations*. arXiv (2018) https://doi.org/10.48550/arXiv.1802.05365

[135] Platt, J.: *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Advances in Kernel Methods-Support Vector Learning, 208 (1998)

[136] Presutti, V., Draicchio, F., Gangemi, A.: *Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames*. Knowledge Engineering and Knowledge Management, 7603, pp. 114–129 (2012) https://doi.org/10.1007/978-3-642-33876-2_12

[137] Presutti, V., Consoli, S., Nuzzolese, A. G., et al.: *Uncovering the Semantics of Wikipedia Pagelinks*. EKAW, 8876 (2014) https://doi.org/10.1007/978-3-319-13704-9_32

[138] Qiu, X., Sun, T., Xu, Y. et al.: *Pre-trained models for natural language processing: a survey*. Science China Technological Sciences, 63, pp. 1872–1897 (2020) https://doi.org/10.1007/s11431-020-1647-3

[139] Radford, A., Narasimhan, K., Salimans, T., et al.: *Improving Language Understanding by Generative Pre-Training*. Technical report, OpenAI (2018)

[140] Reforgiato Recupero, D., Presutti, V., Consoli, S., et al.: *Sentilo: Frame-based sentiment analysis*. Cognitive Computation, 7, pp. 211-225 (2014) https://doi.org/10.1007/s12559-014-9302-z

[141] Reimers, N., Gurevych, I.: *Sentence-bert: Sentence embeddings using siamese bert-networks*. arXiv (2019) https://doi.org/10.48550/arXiv.1908.10084

[142] Reyes, A., Rosso, P., Veale, T.: *A multidimensional approach for detecting irony in Twitter*. Language Resources & Evaluation 47, pp. 239–268 (2013) https://doi.org/10.1007/s10579-012-9196-x

[143] Riboni D., Bettini, C.: *Context-aware activity recognition through a combination of ontological and statistical reasoning*. International Conference on Ubiquitous Intelligent Computing, pp. 39-53 (2009) https://doi.org/10.1007/978-3-642-02830-4_5

[144] Roventini A., Alonge A., Calzolari N., et al.: *ItalWordNet: a Large Semantic Database for Italian*. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 2, Paris, The European Language Resources Association (ELRA), pp. 783-790 (2000)

[145] Ruppenhofer, J., Ellsworth, M., Petruck, M., et al.: *FrameNet II: Extended Theory and Practice*, Jan. (2010)

[146] Sánchez, D., Batet, M., Viejo, A. *Automatic general-purpose sanitization of textual documents*. IEEE Transactions on Information Forensics and Security, 8, 6, pp. 853-862 (2013) https://doi.org/10.1109/TIFS.2013.2239641

[147] Sánchez, D., Batet, M., Viejo, A. *Utility-preserving sanitization of semantically correlated terms in textual documents*. Information Sciences, 279, pp. 77-93 (2014) https://doi.org/10.1016/j.ins.2014.03.103

[148] Sánchez, D., Batet, M.: *C-sanitized: A privacy model for document redaction and sanitization: C-Sanitized: A Privacy Model for Document Redaction and Sanitization*. Journal of the Association for Information Science and Technology, 67 (2016) https://doi.org/10.1002/asi.23363

[149] Santos, D., Cardoso, N.: *A golden resource for named entity recognition in Portuguese*. International Workshop on Computational, Processing of the Portuguese Language, Springer: Berlin/Heidelberg, Germany, 25, pp. 69–79 (2006) https://doi.org/10.1007/11751984_8

[150] Schild, U.J., *The Use of Meta-Rules in Rule Based Legal Computer Systems*. Proceedings of the Fourth International Conference on AI and Law, ACM Press: New York, pp. 100-109 (1993) https://doi.org/10.1145/158976.158989

[151] Schneider, E.W.: *Course Modularization Applied: The Interface System and Its Implications For Sequence Control and Data Analysis*. Association for the Development of Instructional Systems (ADIS), Chicago, Illinois, April 1972 (1973)

[152] Schuster, M., Paliwal, K. K.: *Bidirectional recurrent neural networks*. IEEE Transactions on Signal Processing, 45, pp. 2673–2681 (1997) https://doi.org/10.1109/78.650093

[153] Shannon, C.E.: *A mathematical theory of communication*. The Bell System Technical Journal, 27, 3, pp. 379-423 (1948)

[154] Shapiro, S.: *Vagueness, Open-Texture, and Retrievability*. Inquiry: An Interdisciplinary Journal of Philosophy, 56, 2-3, pp. 307-326 (2013) https://doi.org/10.1080/0020174X.2013.784486

[155] Socher, R., Huang, E.H, Pennington, J. et al.: *Dynamic pooling and unfolding recursive autoencoders for paraphrase detection*. Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11), Curran Associates Inc., Red Hook, NY, USA, pp. 801–809 (2011)

[156] Sokolova, M., Emam, K.: *Evaluation of Learning from Screened Positive Examples*. Proceedings of the 3rd Workshop on Evaluation Methods for Machine Learning, in conjunction with the 25th International Conference on Machine Learning (ICML2008), Helsinki, Finland (2008)

[157] Solove, D. J.: *Understanding Privacy*. Harvard University Press, GWU Legal Studies Research Paper No. 420 (2008)

[158] Sovrano, F., Palmirani, M., Vitali, F.: *Legal Knowledge Extraction for Knowledge Graph Based Question-Answering*. Legal Knowledge and Information Systems JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9–11, pp. 143-153 (2020)

[159] Spinello, R. A.: *Privacy and social networking technology.* The International Review of Information Ethics, 16, pp. 41-60 (2011)

[160] Stoeckli, E., Dremel, C., Uebernickel, F. et al.: *How affordances of chatbots cross the chasm between social and traditional enterprise systems.* Electronic Markets, 30(2), pp. 369-403 (2020) https://doi.org/10.1007/s12525-019-00359-6

[161] Suchomel, V.Pomikálek, J.: *Efficient Web Crawling for large Text Corpora.* Proceedings of the Seventh Web as Corpus Workshop, pp. 1-5 (2012)

[162] Šuster, S., Tulkens, S., Daelemans, W.; *A short review of ethical challenges in clinical natural language processing.* Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain. Association for Computational Linguistics, pp. 80–87 (2017) https://doi.org/10.18653/v1/W17-1610

[163] Swayamdipta, S., Thomson, S., Dyer, C. et al.: *Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold.* arXiv (2017) https://doi.org/10.48550/arXiv.1706.09528

[164] Sweeney, L: *K-Anonymity: a Model for Protecting Privacy.* International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems. 10, 5, pp. 557-570 (2002) https://doi.org/10.1142/S0218488502001648

[165] Tamašauskaitė, G., Groth, P.: *Defining a Knowledge Graph Development Process Through a Systematic Review.* ACM Trans. Softw. Eng. Methodol. (2022) https://doi.org/10.1145/3522586

[166] Tan, S., Na, J: *Positional attention-based frame identification with BERT: A deep learning approach to target disambiguation and semantic frame selection.* arXiv (2019) https://doi.org/10.48550/arXiv.1910.14549

[167] Timmer, R., Liebowitz, D., Nepal S., Kanhere, S.: *Can pre-trained Transformers be used in detecting complex sensitive sentences? - A Monsanto case study.* Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, GA, USA, pp. 90-97 (2021) https://doi.org/10.1109/TPSISA52974.2021.00010

[168] Tomlinson, S.: *Learning task experiments in the trec 2010 legal track.* Proceedings of the 19rd Text REtrieval Conference, TREC '10 (2010)

[169] Torre, D., Abualhaija, S., Sabetzadeh, M. et al.: *An AI-assisted approach for checking the completeness of privacy policies against GDPR.* Proc. IEEE 28th Int. Requirements Eng. Conf. (RE), pp. 136–146 (2020) https://doi.org/10.1109/RE48521.2020.00025

[170] Uschold, M., Grüninger, M.: *Ontologies: Principles, methods and applications.* The Knowledge Engineering Review, 11 (1996) https://doi.org/10.1017/S0269888900007797

[171] Van Der Maaten, L., Hinton, G.: *Visualizing data using t-sne.* Journal of Machine Learning Research, 9, 86, pp. 2579–2605 (2008)

[172] Van Heijst, G.: *The Role of Ontologies in Knowledge Engineering.* Ph.D. Dissertation, University of Amsterdam (1995)

[173] Vasalou, A., Gill, A., Mazanderani, F. et al.: *The Prototype of Privacy: Analysing Privacy Discourse Through its Features* (2010)

[174] Vasalou, A., Gill, A., Mazanderani, F. et al.: *Privacy Dictionary: A New Resource for the Automated Content Analysis of Privacy.* Journal of the American Society for Information Science and Technology, 62, pp. 2095-2105 (2011) https://doi.org/10.1002/asi.21610

[175] Vaswani, A., Shazeer, N., Parmar, N., et al.: *Attention is All you Need.* Guyon, I., and Von Luxburg, V., Bengio S., et al.: Advances in Neural Information Processing Systems, Curran Associates, Inc., 30 (2017)

[176] Visser, P.R.S., Bench-Capon, T.J.M., *A comparison of four ontologies for the design of legal information systems*, Artificial Intelligence and Law, 6, 27-57 (1997) https://doi.org/10.1023/A:1008251913710

[177] Waismann, F.: *Verifiability.* Proceedings of the Aristotelian Society, Supplementary Volume 19, 119–150; reprinted in Logic and language, ed. Antony Flew. Oxford: Basil Blackwell, 1968, pp. 117–144 (1947)

[178] Wang, L., Zhao, X.: *Improved knn Classification Algortihm Research in Text Categorization.* Proceedings of the 2nd International Conference on Communications and Networks (CECNet), pp. 1848-1852 (2012) https://doi.org/10.1109/CECNet.2012.6201850

[179] Weidinger, L., Mellor, J.F.J., Rauh, M. et al.: *Ethical and social risks of harm from language models.* arXiv (2021) https://doi.org/10.48550/arXiv.2112.04359

[180] Westin, A. F.: *Privacy and freedom.* New York, Atheneum (1967)

[181] Wittgenstein, L.: *Philosophical Investigations.* Blackwell Publishing, Oxford (1953)

[182] Xu G., Qi C., Yu H. et al.: *Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network.* International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 474-479 (2019) https://doi.org/10.1109/CyberC.2019.00087

[183] Yu, E., Han, W., Tian, Y., et al.: *ToHRE: A Top-Down Classification Strategy with Hierarchical Bag Representation for Distantly Supervised Relation Extraction.* Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), International Committee on Computational Linguistics, pp. 1665–1676 (2020) https://doi.org/10.18653/v1/2020.coling-main.146

[184] Yue, X., Du, M., Wang, T. et al.: *Differential privacy for text analytics via natural text sanitization.* Proceedings Findings Association Computing Linguistics, pp. 3853–3866 (2021)

[185] Zellig Harris, S.: *Distributional structure.* WORD, 10(2-3), pp. 146–162 (1954) https://doi.org/10.1007/978-94-009-8467-7_1

[186] Zheng, G., Mukherjee, S., Dong, X. L. et al.: *OpenTag: Open Attribute Value Extraction from Product Profiles.* Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, pp. 1049–1058 (2018) https://doi.org/10.1145/3219819.3219839

[187] Baum, Hedlund, Aristei, Goldman. Monsanto papers | secret documents, https://www.wisnerbaum.com/toxic-tort-law/monsanto-roundup-lawsuit/monsanto-papers/. Last access January 18, 2023

[188] Bonatti, P.A., Kirrane, S., Petrova, I.M. et al. The SPECIAL Usage Policy Language, http://purl.org/specialprivacy/policylanguage. Last access January 18, 2023

[189] Cost of a Data Breach Report, https://www.ibm.com/downloads/cas/RDEQK07R. Last access January 18, 2023

[190] General Data Protection Regulation (GDPR), https://www.privacy-regulation.eu/en/4.htm. Last access January 18, 2023

[191] Google De-identify Sensitive Data tool, https://cloud.google.com/dlp/docs/deidentify-sensitive-data#api_overview. Last access January 18, 2023

[192] IBM Discover Sensitive Data tool, https://www.ibm.com/docs/en/guardium/10.6?topic=discover-sensitive-data. Last access January 18, 2023

[193] Microsoft PII detection tool, Azure cognitive service, https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/personally-identifiable-information/overview. Last access January 18, 2023

[194] W3C Data Privacy Vocabulary, https://w3c.github.io/dpv/dpv/. Last access January 18, 2023

Appendix A

Sensitive Information Detection

Table A.1 resumes related work specifying the following aspects: [Work] indicates the analyzed work; [Domain] specifies the type of sensitive information analyzed; [Lang.] indicates the language in which the study is conducted; [Context] describes the span taken into consideration for identification; [CA] (Context-Awareness) indicates if the identification follows a context-aware approach; [Approach] specifies the models and techniques used; [Results] reports the results of experiments if any.

Work	Domain	Lang.	Context	CA	Approach	Results
Battaglia et al. 2020 [12]	PD	English	Post-level	X	BoW and word embeddings	68% F1
Chow et al. 2008 [26]	Enterprise docs	English	Word-level	X	Inference model based on association rule mining	81% rec.; 73% prec.
Dias et al. 2020 [35]	Basic PD	Portug.	Entity level	V	Rules-based, ML and DL methods	CRF 65.50%; Bi-LSTM 83.01% F1
Garcia et al. 2017 [54]	Enterprise docs	English	Paragraph and doc-level	X	Inference ontological model	/
Genetu et al. 2021 [57]	PD (health, political, ethnic)	Amharic	Sentence-level	V	LSTM, CNNs, Bi-LSTM	Detect. 90% acc.; classif. 93% acc.
Geng et al. 2011 [58]	Medical entities	English	Word-level	X	Inference ontological model	84% rec.; 74% prec.; 79% F1
Guo et al. 2021 [64]	PD	English	Sentence-level	X	BERT Bi-LSTM	99.5% F1
Hart et al. 2011 [68]	Enterprise docs	English	Doc-level	X	SVMs based on unigram features	FDR 0.46%; FNR 1.6%

Islam et al. 2014 [78]	Social network private behavior	English	User-level	X	AdaBoost with NB classifier and LDA	Binary task 95.45% acc.; 3-class task 69.63%
Korba et al. 2008 [87]	PII	English	Word-level	X	Rules-based and ML NER and RE	93.3% acc.
Macdonald et al. 2017 [104]	Gov. docs	English	Doc-level	X	Word embeddings	0.54% F2
Neerbek et al. 2020 [121]	Enterprise docs	English	Sentence-level	V	Paraphrase using RecNN	Silver dataset 83% acc.; golden dataset 81,7% acc.
Obeid et al. 2019 [124]	Clinical text	English	Doc-level	V	Word embedding based CNNs	95% acc.
Pablos et al. 2020 [55]	Medical docs	Spanish	Sentence-level	V	BERT model	MEDOCCAN 97% F1; NUBES-PHI 95% F1
Sanchez et al. 2014a,b [148, 147]	Unstruct. docs	English	Word-level	X	PMI inference model	Average 97% of sanitization docs
Sokolova et al. 2008 [156]	Medical docs	English	Word-level	X	Rules-based NER	/
Timmer et al. 2021 [167]	Entreprise docs	English	Sentence-level	V	BERT model	Silver dataset 84% acc.
Xu et al. 2019 [182]	Military and politically docs	Chinese	Sentence-level	V	Text-CNN	95.17% acc.

Table A.1: Review of sensitive corpora identification related work

Appendix B

Privacy Corpora

Corpus	Date	Domain	Dimension	Silver-labels	Golden-labels	Related work
Enron corpus	2003	Company email dataset	619,446 emails, 158 users	X	V	[26, 68, 120, 91]
Monsanto corpus	2017	Legal domain	274 docs	V	V	[121, 167]
WikiLeaks dataset	2011	Military and religious organization	DynCorp (23 docs); TM (120 docs); Mormon corpus	V	X	[68]
Wikipedia dataset	2011 (1), 2013 (2)	Sensitive articles and pages (1), PII key-word from wiki pages	10k articles (1)	X(1,2)	V(1,2)	[68, 148, 146, 147]
Google private dataset	2011	Posts of soft-develop blogs	1119 posts	V	X	[68]
Government records dataset	2007	Personal information and international relations	1,111 records	X	V	[14, 108]
Twitter corpus	2014	User behavior and content	426,464 tweets	X	V	[78]
Pastebin dataset	2021	PII, credential and financial inf.	1,035,634 docs	V	V	[64]
Social media posts dataset	2020	Sensitive and non-sensitive posts	679 posts	X	V	[12]

Table B.1: Review of sensitive corpora in related work

Appendix C

Complete Feasibility Analysis PDCs from DPV

Feasibility type	**PDCs**
Macro-categories	INTERNAL, EXTERNAL, SOCIAL, FINANCIAL, TRACKING, HISTORICAL
Identifiable through textual elements	AGE, AGE EXACT, AGE RANGE, APARTMENT OWNED, BIRTH DATE, BIRTH PLACE, CAR OWNED, COUNTRY, CREDIT, CRIMINAL, CRIMINAL CHARGE, CRIMINAL CONVICTION, CRIMINAL PARDON, CRIMINAL OFFENSE, CURRENT EMPLOYMENT, DEMOGRAPHIC, DIALECT, DISABILITY, DISCIPLINARY ACTION, DIVORCE, DRUG TEST RESULT, EDUCATION, EDUCATION EXPERIENCE, EDUCATION QUALIFICATION, EMPLOYMENT HISTORY, ETHNICITY, ETHNIC ORIGIN, FAMILY, FAMILY HEALTH HISTORY, FAMILY STRUCTURE, FAVORITE, FAVORITE COLOR, FAVORITE FOOD, FAVORITE MUSIC, FETISH, GENDER, GEOGRAPHIC, HAIR COLOR, HEALTH, HEALTH HISTORY, HEALTH RECORD, HEIGHT, HOUSE OWNED, INCOME BRACKET, INDIVIDUAL HEALTH HISTORY, JOB, LANGUAGE, LOCATION, MARITAL STATUS, MARRIAGE, MEDICAL HEALTH, MENTAL HEALTH, NAME, NATIONALITY, OFFSPRING, OWNERSHIP, PARENT, PAST EMPLOYMENT, PERFORMANCE AT WORK, PERSONAL DOCUMENTS, PERSONAL POSSESSION, PHYSICAL CHARACTERISTIC, PHYSICAL HEALTH, PHYSICAL TRAIT, PIERCING, POLITICAL AFFILIATION, POLITICAL OPINION, PRESCRIPTION, PRIVACY PREFERENCE, PROCLIVITIE, PROFESSIONAL, PROFESSIONAL CERTIFICATION, PROFESSIONAL EVALUATION, PROFESSIONAL INTERVIEW, RACE, REFERENCE, RELATIONSHIP, RELIGION, SALARY, SCHOOL, SEXUAL, SEXUAL HISTORY, SEXUAL PREFERENCE, SIBLING, SKIN TONE, TATTOO, TRAVEL HISTORY, WEIGHT, WORK HISTORY, WORK ENVIRONMENT
Broad-boundaries categories	ATTITUDE, BEHAVIORAL, CHARACTER, COMMUNICATION, COMMUNICATION METADATA, DEMEANOR, DISLIKE, GENERAL REPUTATION, HOUSEHOLD DATA, IDENTIFIER, IDENTIFYING, INTENTION, INTERACTION, INTEREST, KNOWLEDGE BELIEF, LIFE HISTORY, LIKE, OPINION, PERSONALITY, PHILOSOPHICAL BELIEF, PREFERENCE, PUBLIC LIFE, RELIABILITY, RELIGIOUS BELIEF, SOCIAL STATUS, THOUGHT
Uniquely identifiable	ACCOUNT IDENTIFIER, AUTHENTICATING, BANK ACCOUNT, BLOOD TYPE, CONTACT, CREDIT CARD NUMBER, CREDIT SCORE, EMAIL ADDRESS, EMAIL ADDRESS PERSONAL, EMAIL ADDRESS WORK, FINANCIAL ACCOUNT, FINANCIAL ACCOUNT NUMBER, FINANCIAL STATUS, INSURANCE, IPAddress, MACAddress, OFFICIAL ID, PASSPORT, PIN CODE, PASSWORD, PHYSICAL ADDRESS, ROOM NUMBER, SECRET TEXT, TELEPHONE NUMBER, UID, USERNAME, VEHICLE LICENSE NUMBER, VEHICLE LICENSE DATA, VEHICLE LICENSE, VEHICLE USAGE DATA
Identifiable through non-textual elements	ACCENT, ACQUAINTANCE, ASSOCIATION, AUTHENTICATION HISTORY, BIOMETRIC, BROWSING BEHAVIOR, BROWSER FINGERPRINT, BROWSER HISTORY, BROWSING REFERRAL, CALL LOG, CONNECTION, CREDIT, CREDIT CAPACITY, CREDIT RECORD, CREDIT STANDING, CREDIT WORTHINESS, DEVICE APPLICATIONS, DEVICE-BASED, DIGITAL FINGERPRINT, DNA CODE, DEVICE OPERATING SYSTEM, DEVICE SOFTWARE, EMAIL CONTENT, FACIAL PRINT, FINGERPRINT, FRIEND, GENETIC DATA, GPS COORDINATE, GROUP MEMBERSHIP, INCOME, LINK CLICKED, LOAN RECORD, PAYMENT CARD, PAYMENT CARD EXPIRING, PAYMENT CARD NUMBER, PICTURE, PROFILE, PUBLICLY AVAILABLE SOCIAL MEDIA DATA, PURCHASE, PURCHASE AND SPENDING HABITS, RETINA, SALE, SERVICE CONSUMPTION BEHAVIOR, SOCIAL MEDIA COMMUNICATION, SOCIAL MEDIA DATA, SOCIAL NETWORK, TAX, TRADE UNION MEMBERSHIP, TRANSACTION, TRANSACTIONAL, TV VIEWING BEHAVIOR, USER AGENT, VOICE COMMUNICATION RECORDING, VOICE MAIL

Table C.2: Listed PDCs from DPV and feasibility subdivision

Appendix D

SPEDAC 3 PDCs

Label	PDCs
[Age]	AGE, AGE EXACT, AGE RANGE, BIRTH DATE, BIRTH PLACE
[Apartment Owned]	APARTMENT OWNED
[CarOwned]	CAR OWNED
[Country]	COUNTRY
[Credit]	CREDIT
[Criminal]	CRIMINAL, CRIMINAL CHARGE, CRIMINAL CONVICTION, CRIMINAL PARDON, CRIMINAL OFFENSE
[Dialect]	DIALECT
[Disability]	DISABILITY
[Divorce]	DIVORCE
[Drug Test Result]	DRUG TEST RESULT
[Employment History]	EMPLOYMENT HISTORY
[Ethnicity and Ethnic Origin]	ETHNICITY, NATIONALITY
[Family and Family Structure]	FAMILY, FAMILY STRUCTURE
[Family Health History]	FAMILY HEALTH HISTORY
[Favorite]	FAVORITE
[Favorite Color]	FAVORITE COLOR
[Favorite Food]	FAVORITE FOOD
[Favorite Music]	FAVORITE MUSIC
[Fetish]	FETISH
[Gender]	GENDER
[Hair Color]	HAIR COLOR
[Health]	HEALTH, HEALTH RECORD, MEDICAL HEALTH
[Health History]	HEALTH HISTORY, INDIVIDUAL HEALTH HISTORY
[Height]	HEIGHT
[House Owned]	HOUSE OWNED
[Income Bracket]	INCOME BRACKET
[Job]	JOB
[Language]	LANGUAGE
[Life History]	LIFE HISTORY
[Location]	LOCATION, GEOGRAPHIC, DEMOGRAPHIC

[Marital Status]	MARITAL STATUS
[Marriage]	MARRIAGE
[Mental Health]	MENTAL HEALTH
[Name]	NAME
[Offspring]	OFFSPRING
[Ownership]	OWNERSHIP, PERSONAL POSSESSION, PERSONAL DOCUMENTS
[Parent]	PARENT
[Physical Characteristic and Trait]	PHYSICAL CHARACTERISTIC, PHYSICAL TRAIT
[Physical Health]	PHYSICAL HEALTH
[Piercing]	PIERCING
[Political Affiliation]	POLITICAL AFFILIATION, POLITICAL OPINION
[Prescription]	PRESCRIPTION
[Privacy Preference]	PRIVACY PREFERENCE
[Proclivitie]	PROCLIVITIE
[Professional]	PROFESSIONAL, CURRENT EMP., PAST EMP., WORK ENVIRONMENT
[Professional Certification]	PROFESSIONAL CERTIFICATION
[Professional Evaluation]	PROFESSIONAL EVALUATION, PERFORMANCE AT WORK, DISCIPLINARY ACTION
[Professional Interview]	PROFESSIONAL INTERVIEW
[Race]	RACE
[Reference]	REFERENCE
[Relationship]	RELATIONSHIP
[Religion]	RELIGION
[Salary]	SALARY
[School]	SCHOOL, EDUCATION, EDUCATION EXPERIENCE, EDUCATION QUALIFICATION
[Sexual]	SEXUAL
[Sexual History]	SEXUAL HISTORY
[Sexual Preference]	SEXUAL PREFERENCE
[Sibling]	SIBLING
[Skin Tone]	SKIN TONE
[Tattoo]	TATTOO
[Weight]	WEIGHT
[Work History]	WORK HISTORY

Table D.1: Listed of labels in SPEDAC 3 and PDCs included

Appendix E

PRIVAFRAME Broad-boundaries PDCs

Broad-boundaries PDCs represented in PRIVAFRAME
[Attitude]
[Behavioral]
[Character]
[Communication]
[Communication Metadata]
[Demeanor]
[Dislike]
[General Reputation]
[Identifying]
[Intention]
[Interaction]
[Interest]
[Knowledge Belief]
[Opinion]
[Personality]
[Preference]
[Religious Belief]
[Social Status]
[Thought]

Table E.1: List of broad-boundaries PDCs .