Computers in Biology and Medicine 185 (2025) 109576

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Named entity recognition for de-identifying Spanish electronic health records

Francisco J. Moreno-Barea [a,*], Guillermo López-García [b], Héctor Mesa [a], Nuria Ribelles [c], Emilio Alba [c], José M. Jerez [a,d], Francisco J. Veredas [a,d]

[a] Departamento de Lenguajes y Ciencias de la Computación, Escuela Técnica Superior de Ingeniería Informática, Universidad de Málaga, Málaga, Spain
[b] Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, USA
[c] Unidad de Gestión Clínica Intercentros de Oncología, Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain
[d] Research Institute of Multilingual Language Technologies, Universidad de Málaga, Málaga, Spain

ARTICLE INFO

Keywords:
Named entity recognition
Natural language processing
De-identification
Electronic health records
Spanish

ABSTRACT

Background and objectives: There is an increasing and renewed interest in Electronic Health Records (EHRs) as a substantial information source for clinical decision making. Consequently, automatic de-identification of EHRs is an indispensable task, since their dissociation from personal data is a necessary prerequisite for their dissemination. Nevertheless, the bulk of prior research in this domain has been conducted using English EHRs, given the limited availability of annotated corpora in other languages, including Spanish.
Methods: In this study, the automatic de-identification of medical documents in Spanish was explored. A private corpus comprising 599 genuine clinical cases was annotated with eight different categories of protected health information. The prediction problem was approached as a named entity recognition task and two deep learning-based methodologies were developed. The first strategy was based on recurrent neural networks (RNN) and the second, an end-to-end approach, was based on Transformers. In addition, we have implemented a procedure to expand the amount of texts employed for model training.
Results: Our findings demonstrate that Transformers surpass RNNs in the de-identification of clinical data in Spanish. Particularly noteworthy is the excellent performance of the XLM-RoBERTa large Transformer, achieving a rigorous strict-match micro-average of 0.946 for precision, 0.954 for recall, and an F1 score of 0.95 when applied to the amplified version of the corpus. Furthermore, a web-based application has been created to assist specialized clinicians in de-identifying EHRs through the aid of the implemented models.
Conclusion: The study's conclusions showcase the practical applicability of the state-of-the-art Transformers models for precise de-identification of clinical notes in real-world medical settings in Spanish, with the potential to improve performance if continual pre-training strategies are implemented.

1. Introduction

1.1. EHRs and de-identification

The widespread adoption of electronic health records (EHRs) has emerged as a pivotal cornerstone within healthcare systems, essential for both medical professionals and researchers. EHRs serve as invaluable reservoirs of information, offering unprecedented potential to advance medical research and enhance healthcare services. Nevertheless, leveraging EHRs for broader medical research purposes necessitates the resolution of certain inherent challenges, with paramount among them being the heterogeneous nature of the contained information and the imperative need to safeguard patient privacy. These challenges are inherently intertwined.

EHRs house a diverse array of data, ranging from structured data to unstructured free-text documents, such as clinical notes and radiology reports. These documents encapsulate a wealth of clinical information encompassing diagnoses, treatments, procedures, and, significantly, the privacy-sensitive details of patients and healthcare providers. However, the unstructured nature of these textual components presents formidable obstacles to the automated extraction of pertinent clinical concepts. Manual extraction, though feasible, is neither scalable nor cost-effective [1,2].

The fundamental aspect of utilizing EHRs for subsequent medical-analytical processes lies in extracting concepts linked to personally

* Corresponding author.
E-mail address: fjmoreno@lcc.uma.es (F.J. Moreno-Barea).

https://doi.org/10.1016/j.compbiomed.2024.109576
Received 3 January 2024; Received in revised form 31 October 2024; Accepted 11 December 2024
Available online 21 December 2024

identifiable health information. De-identification, which is a crucial task, involves recognizing and eradicating protected health information (PHI) to maintain the inviolability of patient privacy. This requirement, guided by ethical considerations and enforced by legal obligations, is demonstrated by the Health Insurance Portability and Accountability Act (HIPAA) in the United States [3]. HIPAA mandates the elimination of 18 categories of protected health information (PHI) [4]. Similarly, the General Data Protection Regulation (GDPR) of the European Union (EU) [5] and the *Ley Orgánica Española de Protección de Datos Personales y Garantía de Derechos Digitales (LOPD-GDD)* of Spain [6] strictly forbid the processing of personal data without prior de-identification [6].

Taking into consideration the regulations and terminology used, it is necessary to differentiate between de-identification and anonymization. De-identification involves concealing or removing explicit identifiers, while anonymization ensures that the data cannot be traced back to identify the patient. Essentially, de-identification may not render the data entirely anonymous [7]. Automated de-identification typically involves three fundamental tasks: identifying references to personal information in electronic health records, classifying personal information into predetermined categories, and replacing these references with anonymized labels or realistic surrogates. This study is centered on de-identifying Spanish electronic health records to comply with the strict regulations of the LOPD-GDD.

From the perspective of natural language processing (NLP), the de-identification of clinical texts closely aligns with Named Entity Recognition (NER). NER entails detecting text segments that refer to rigid concepts belonging to predefined semantic types, such as persons, locations, and organizations [8]. NER is a crucial preprocessing step for various NLP applications. It has significant importance in NLP since it excludes subjective evaluations, ensures logical structures, and maintains formal writing styles [9–13]. NER assists in information retrieval, machine translation, automatic summarization, knowledge base construction, and question answering. The concept of Named Entities (NEs) rose to prominence at the 6th Message Understanding Conference (MUC6) [14], a seminal gathering focused on fostering and evaluating research in information extraction. Initially aimed at identifying names of individuals, organizations, locations, and numerical expressions such as dates and times, the scope of NEs now extends to identifying protected health information in EHR de-identification processes.

1.2. Text de-identification approaches

Over the past few decades, the process of text de-identification has significantly evolved and now encompasses three distinct approaches: rule-based, machine learning-based (ML), and deep learning-based (DL) methodologies. In the clinical domain, during the initial phases of de-identification systems, rule-based strategies were mainly implemented [15–17]. These systems relied heavily on manually created rules, patterns, and specialized semantic dictionaries to identify relevant entities in textual information. However, it is crucial to recognize the limitations of such rule-based approaches, as they are not easily transferable between systems. Rules tailored for one system can often prove difficult to adapt to another.

Due to the limitations of rule-based methods, the research community has turned to the creation of ML algorithms. This was prompted by the organization of de-identification challenges, including the 2006 i2b2 NLP challenge, the 2014 i2b2 NLP challenge, and the 2016 CEGS N-GRID NLP challenge. The employed ML algorithms for this task comprised decision trees (DT) [18], hidden Markov models (HMM) [19], support vector machines (SVM) [20], structured support vector machines (SSVM) [21], and conditional random fields (CRF) [22]. It is worth noting that for the task of de-identification, CRF, SSVM, and HMM framed the NER task as a sequential labeling problem. This involved assigning a class label to each word in a sequence, and exploiting contextual interactions between neighboring labels. Among these, CRF is the leading state-of-the-art (SOTA) method evidenced by its prominence in the primary systems of the 2014 i2b2 de-identification challenge.

In recent years, deep neural networks (DNN), notable for their ability to learn effective features independently from vast unlabeled data, have become a critical component in various NLP tasks. As a result of this development, there is no more need for intricate feature engineering, which is a characteristic hallmark of traditional statistical learning approaches. A variety of adapted and combined DNN structures have been utilized for marking tasks, stretching from basic feed-forward neural networks to structures based on recurrent neural networks (RNN). Various neural network architectures have commonly been employed for NER, including the Gated Recurrent Unit (GRU) [23], Long Short-Term Memory (LSTM) networks [24], and various combinations such as LSTM-CRF [25], LSTM-CNNs [26], and CNN-LSTM-CRF [27]. Notably, the character-level bidirectional LSTM-CRF structure [28] has garnered attention for yielding SOTA NER performance.

Presently, the landscape of NLP is reshaped by large pretrained language models supported by multi-head self-attention and based on the Transformer architectures [29], with Bidirectional Encoder Representations from Transformers (BERT) [30] leading the way. These models have surpassed traditional ML systems, particularly in the domain of NER, and have gained considerable prominence in the biomedical field [31,32]. The ability to identify important elements exceeds past standards, signaling a significant advancement in both text anonymization and biomedical informatics.

1.3. Contributions and structure

Considering all the above aspects, in this paper we have addressed the problem of automatic detection of personally identifiable information in Real-World Data (RWD) from EHR written in Spanish. For this purpose, we have used several models based on RNN (BiLSTM, 2-BiLSTM and BiLSTM-CRF) and the Transformer architecture. We have experimented with models trained with multilingual corpora (XLM-R [33], XLM-R-Galén [34]) and others with monolingual corpora in Spanish (RoBERTa-Bio [35], RoBERTa-BNE [36]). In addition, some of them are trained or tuned with general purpose corpora (XLM-R, RoBERTa-BNE) and others trained or tuned to adapt them to the biomedical or clinical domain (RoBERTa-Bio, XLM-R-Galén). The models studied herein represent the SOTA in various NER tasks [37].

On a previous work [38], we conducted a preliminary study in this regard. The results presented in that paper included a comparative analysis of the performance of three different RNNs and two Transformer-based models (XLM-RoBERTa and RoBERTa-BNE) when dealing with a NER de-identification task carried on a section of our proprietary EHR corpus Galén [39]. Unlike the present study, we did not conduct multiple-run tests or experimented with other medical-domain corpora in Spanish. This preliminary work provided a less detailed performance analysis, less confidence in the conclusions yielded and it did not incorporate transfer-learning or continuous pretraining strategies. So, apart from our previous work, to the best of our knowledge, this is the first study that analyses the application of these models to the problem of identifying PHI to clinical corpora with real-world data in Spanish, including experimentation on the models' ability to handle other public Spanish corpus used in competitive tasks, such as the MEDDOCAN corpus [40].

Through this study, we show how the Transformers applied to the task of de-identification of clinical texts in Spanish achieve higher performance rates than RNN-based models, thus supporting the fact that Transformers represent currently the SOTA for many NER tasks. Additionally, we introduce herein a web application that has been designed to aid medical and clinical professionals in tasks of de-identifying EHRs. The software's core has been supplied with the de-identification models that obtained the best results in our analysis. As part of its operation, this tool enables users to upload and edit documents, launch

the de-identification process, correct minor errors made by the NER models and obtain de-identified texts as results, through different transformation profiles aimed at minimizing the probability of document re-identification.

The paper is structured as follows: Section 2 shows some NER challenges related works. Section 2.1 introduces the materials, the Galén corpus, the labeling process and the selected NEs. Section 3 introduces the methodology, the strategies designed for obtaining token embeddings and, finally, the RNN and Transformers models employed in this study. Section 4 presents the results of this study and is structured into three subsections: a first one with the description of the experiments and the evaluation metrics used, a second one with the comparison analysis of the performance results obtained with the different NER models studied, and a last one with the results of a detailed evaluation at the NE-level. Section 5 introduces the publicly available version of the Spanish clinical text de-identification tool developed for this work, with an emphasis on its interface and functionality. We finalize our paper with a discussion in Section 6 and relevant conclusions in Section 7.

2. Related works

With the emergence of various NER challenges, including the 2006 i2b2 NLP challenge, the 2014 i2b2 NLP challenge, and the 2016 CEGS N-GRID NLP challenge, and the widespread adoption of EHRs in healthcare systems globally, studies on text de-identification have flourished. The 2006 and 2014 challenges demonstrated significant advancements in this domain, with CRF models proving to be the most proficient systems at the time. Thus, the model presented in [41], which utilized word-token, context, orthographic, sentence-level, and dictionary features to train a CRF model, achieved the highest F-measure in the 2014 i2b2 NLP challenge.

More recently, the increasing use of DL has led to the development of architectural models based on RNN. Thus, LSTM networks [24] alone or in combinations with CRF, such as LSTM-CRF [25], LSTM-CNNs [26], and CNN-LSTM-CRF [27], demonstrated superior performance in the 2016 CEGS N-GRID de-identification task when compared to simple CRF models. Dernoncourt et al. (2016) [37] utilized a bidirectional LSTM model that resulted in the SOTA for the 2014 i2b2/UTHealth de-identification challenge. Their proposed strategy integrated a layer of BiLSTM units at a character-level input to obtain character embeddings which are then combined with pretrained token embeddings. These improved token embeddings are returned to the BiLSTM units, and finally the CRF sequence optimizer adjusts the probability sequence to generate the system output. Considering the excellent performance demonstrated by the LSTM-CRF architectures, the majority of the research carried out in the area of medical text de-identification (just from real-world-text or other biomedical-text sources) employs these models [42–45].

Since most de-identification or anonymization techniques have been developed primarily for English corpora, the direct application of the models obtained to other languages may raise not only technical or efficiency problems, but also problems arising from the different data protection legislation prevailing in different countries and regions with official languages other than English. It is for this reason that de-identification models have been developed on the basis of corpora written in different languages, such as French, German, Dutch or Chinese. We can find in the specific literature examples of such models that have followed strategies ranging from the use of dictionaries and specific ontologies to the use of ML strategies [46–49].

Several competitions and projects have endeavored to utilize the data within unstructured Spanish medical records. However, due to the restricted availability of annotated clinical corpora, the task remains challenging. We can start by naming the CANTEMIST challenge, a *CANcer TExt MIning SharedTask* which has produced a number of resources aiming to recognize NEs that are critical in regards to the extraction of concepts of cancer and tumor morphology in Spanish medical records [50,51]. For its part, in the PharmacoNER shared task — which dealt with automatic recognition of chemical and drug mentions from a Spanish corpus consisting of clinical cases — the most effective models proposed relied on contextualized embeddings, like BERT and Flair [52,53]. Lastly, the MEDDOCAN shared task on *MEDical DOCuments ANonimization*, held in 2019 in Bilbao (Spain), targeted the de-identification of clinical texts in Spanish [40]. As shown by the latest NER and de-identification competitions that have been recently held, DL-based methodologies result in better performance compared to traditional rule-based or ML strategies [54,55]. Thus, the winning model in the MEDDOCAN competition was based on BiLSTM-CRF network architecture [54]. Currently, the MEDDOCAN corpus — compiled and adapted from clinical cases extracted from medical articles in Spanish — is employed as a reference dataset to evaluate the performance of de-identification models for medical documents in Spanish, such as radiology reports written in this language [56].

In addition to the aforementioned competitions, other research works have also made significant advances in NER on unstructured medical texts in Spanish. Weegar et al. [57] employed embedding models and RNNs to undertake a NER task with the objective of extracting diseases, drugs, disorders and findings from Swedish and Spanish clinical texts. Santiso et al. [58] employed rule-based and ML techniques to perform negated entity recognition on EHRs in Spanish. This task addresses the recognition of entities and the classification of those that are negated or not. Additionally, experiments conducted on the *n2c2 2022 Track Contextualized Medication Event Extraction* demonstrated that data augmentation with GPT-3 can enhance the performance of a Transformer-based model for NER on the i2b2 2014 Heart Disease Risk Factor Challenge [59].

In the context of other corpora comprising real clinical texts in Spanish, the Chilean Waiting List corpus [60] is worthy of mention. This corpus was conceived with the objective of performing a NER task, which includes the identification of NEs such as procedures, diseases, body parts, medications, abbreviations and family members. The Multiple LSTM-CRF (MLC) achieved SOTA results and was used to develop a software system similar to the one presented in this work [61]. Finally, a similar corpus is being used to label personal information with a lower number of NEs than those proposed in this study, achieving SOTA results with Transformer models [62].

2.1. Materials

The materials, strategies and procedures used to identify, extract and annotate the PHI entities from the different corpora used in this study are described in detail in this section. This includes a description of the Galén corpus used to train and test the models on a real-world clinical scenery, the labeling process, the PHI NEs considered, as well as the data augmentation (DA) procedure performed to optimize the efficacy of the de-identification models are also presented in this section.

2.1.1. Corpora annotation

In general, the design of algorithms for annotation and automatic classification of clinical texts requires manual retrospective analysis of a large amount of unstructured information, for the annotation and labeling of the categories corresponding to the documents used in the design of classification models. In this sense, the availability of the Galén system [39,63], with information on more than 60,000 cancer patients, with a total of 600,000 documents corresponding to clinical episodes and a significant number of structured fields (which are completed both in real time, during clinical care activity, and later by specific personnel in charge of this supervised task), supplies the research team with quality supervised information to address different types of classification and coding problems in the NLP field.

Table 1
Characteristics of the Galén and MEDDOCAN corpus in their original version and with DA x10. The characteristics are included in their absolute value and in terms of their presence per document (P. doc.).

Characteristic	Galén				MEDDOCAN			
	Original		DA x10		Original		DA x10	
	Total	P. doc.	Total	P. doc.	Total	P. doc.	Total	P. doc.
Documents	600		4 641		1 000		8 500	
Paragraphs	29 678	49.46	46 770	10.08	1 000	1.00	8 510	1.00
Sentences	70 283	117.14	141 914	30.58	31 921	31.92	271 808	31.98
Words	723 590	1 205.98	1 674 430	360.79	439 150	43.15	3 752 100	441.42
Diff. Words	33 803	56.34	41 598	8.96	32 742	32.74	56 023	6.59
Non-stop-words	539 924	899.87	1 219 559	262.78	294 992	294.99	2 526 769	297.27
Diff. Non-stop-words	33 446	55.74	41 238	8.89	32 500	32.50	55 773	6.56

Once the information has been extracted, we need to proceed, as it is mandatory according to the LOPD-GDD [6], with the de-identification of the medical records to guarantee their correct anonymization and dissociation of the personal information of the individuals involved in the health system. In the first step, each document generated was referenced by a unique identifier (UUID), eliminating direct associations to numbers or patient identification codes (NUHSA, medical record number or any personal ids). Subsequently, in order to have a subset of labeled records that will be used for the subsequent training of the models, a supervised annotation was made exclusively by authorized clinical personnel of the Intercenter Medical Oncology Clinical Management Unit (UGCOI). Thus, a certain initial volume of clinical records was processed and labeled by our authorized staff, by replacing sensitive information with standardized labels in a non-reversible way. To facilitate this manually labeling task and avoid possible errors once this manual labeling was performed, a preliminary unsupervised annotation of entities (patient identifiers, dates, telephone numbers, e-mail and postal addresses, urls, etc.) was carried out by using classic NLP methods (regular expressions and parsers) based on corpus of medical records published by third parties for training and validation of these algorithms. It was proposed to use this annotation as a starting point for the subsequent manual validation of a subset of EHR texts, which allowed the application of supervised learning techniques for the automatic labeling of these entities, with the consequent saving in time and resources for those responsible for the UGCOI.

On the other hand, the MEDDOCAN corpus [40] was used to test the efficiency of the analyzed de-identification models in different case studies, such as training them with one corpus and testing them on another independent corpus. The MEDDOCAN corpus is composed of clinical cases written in Spanish that belonged to the Spanish National Cancer Research Center (CNIO), which requires that patients' medical records have all personally identifiable data removed to protect patient privacy. However, the main difference with the Galén corpus is that the documents collected at MEDDOCAN are clinical cases published in medical articles with semi-structured information. In other words, MEDDOCAN is not a corpus obtained from real-world medical records. The Galén corpus, on the other hand, belongs to the oncological domain and collects clinical documents written in natural language during clinical practice. In the further exploration of the MEDDOCAN corpus, it compiles to a total of 1000 clinical documents de-identified by staff of the *Hospital Universitario 12 de Octubre*—by applying the guidelines set by HIPAA through double annotation followed by rounds of consistency checking, review and correction, and adjusting to the reality of clinical records in Spain.

Table 1 shows an analysis of the corpora explored in the present study, and their augmented version (DA), which is explained in a later section. The significant characteristics of the corpora, including the number of documents, paragraphs, sentences, words, non-stop words, and different words, are examined. Additionally, the distribution of these elements per document is analyzed. The results demonstrate that the number of sentences and words per document is greater in the Galén corpus than in MEDDOCAN. The average number of paragraphs, sentences, and non-stop words per document in the Galén corpus is 49.5, 117.1, and 899.9, respectively. In comparison, the corresponding characteristics for the MEDDOCAN corpus are 1.0, 31.9, and 295.0. This illustrates the distinction between a synthetic corpus with semi-structured documents, and a real corpus written by clinicians, where the texts are unstructured. It also demonstrates the inherent challenges associated with performing the NER task on the Galén corpus. Additionally, Table 1 illustrates the impact of DA on the corpora, indicating a decrease in the overall average of the features as the number of documents increases, but these being different only in the varied keywords.

Additionally, an analysis has been conducted to measure the presence of non-stop and non-numerical words in the corpora. Fig. 1 shows the distribution of the frequencies of occurrence of non-stop-words in both corpora on a logarithmic scale. It is evident that most of the words appear less than 500 times in both corpora, indicating a considerable number of words that appear infrequently in the documents. In the case of the Galén corpus, it is observed that there are different words that appear more than 3000. In contrast, even though the number of total documents in MEDDOCAN is much higher than in Galén — 600 and 1000 documents, respectively —, the maximum occurrence of a word in the former is around 3000. Furthermore, an increase in the number of words with an occurrence of precisely 1000 can be observed in MEDDOCAN, i.e. exactly one occurrence per document. These words directly reference personal data in the semi-structured documents ('NHC', 'NASS', 'Dirección'/Address, 'Localidad'/Location, 'País'/Country, 'Nombre'/First name, 'Apellidos'/Surname, 'Médico'/Doctor, etc.). This obviously facilitates the identification of NEs by the models on MEDDOCAN. This clearly differs from the distribution observed in Galén, where the frequencies are more distributed, reflecting the unstructured nature of the corpus.

2.1.2. Named entities

Our annotation guidelines are driven by the goals of the national research project of a clinical nature in which our study is immersed. We started by examining the presence of the PHI categories defined by the HIPAA in the US. A reliable interpretation of the HIPAA guidelines was made, adapting some PHIs to fit the reality of health records in Spain. The Spanish legal system does not provide specific guidance on what information must be removed to de-identify medical texts, but the annotation guidelines made by the Spanish National Plan for the Advancement of Language Technology (PlanTL) for the MEDDOCAN shared task [40] were taken into account. The task was also carried out from a position of "risk aversion", due to the great variability of users who would later view and interpret the information based on the de-identification.

A post-selection of PHI categories was made after manually reviewing the data. Table 2 shows the organization of privacy data into 8 final NEs and their presence in the train, validation, and test sets used to the experimentation. The NEs finally chosen were:

- ADDRESS: includes the appearance of a physical address, such as streets, avenues or buildings.

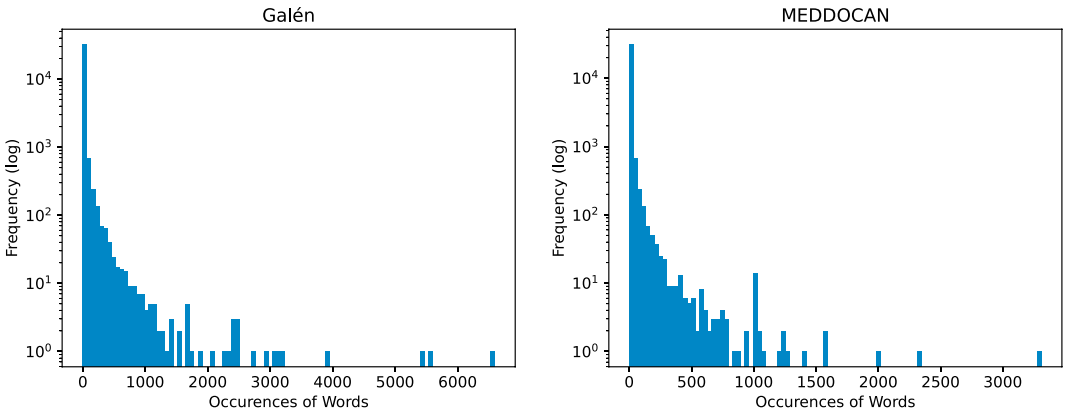


Fig. 1. Distribution of frequencies obtained according to occurrences of non-stop-words in Galén and MEDDOCAN corpora in a log-scale. Note that the horizontal axes of both figures are at different scales.

Table 2
Number (abs) and percentage (%) of annotations per subset in Galén and MEDDOCAN corpora: training, validation and test sets.

NE	Galén corpus						MEDDOCAN corpus					
	Train		Validation		Test		Train		Validation		Test	
	abs	%	abs	%	abs	%	abs	%	abs	%	abs	%
ADDRESS	2	.0013	1	.0026	1	.0029	1701	.2326	854	.2298	829	.2231
CENTER	468	.3145	114	.2953	91	.2600	360	.0492	214	.0576	203	.0546
CONTACT	41	.0276	14	.0363	17	.0486	542	.0741	272	.0732	282	.0759
HISTORY	18	.0121	14	.0363	15	.0429	590	.0807	288	.0775	306	.0824
IDENT	63	.0423	9	.0233	8	.0229	865	.1183	419	.1128	433	.1165
LOCATION	342	.2298	74	.1917	66	.1886	1746	.2387	914	.2460	903	.2430
PERSON	439	.2950	147	.3808	130	.3714	1510	.2065	755	.2032	760	.2045
REFERENCE	115	.0773	13	.0337	22	.0629	0	.0000	0	.0000	0	.0000
Total	1488		386		350		7314		3716		3716	

- CENTER: includes any reference to names of clinical centers, general hospitals, institutions or health centers.
- CONTACT: includes any form of contact with a patient, doctor, nurse or health center, such as telephone numbers and email addresses.
- HISTORY: includes the identifier used to control history numbers, hospital medical records (NHC) and the personal unique number of Andalusian Health History (NUHSA).
- IDENT: includes personal identifiers such as the national identity document (DNI), the social security number (NSS), identifiers associated with insurers, the personal numerical code in the Andalusian Health Service (CNP), or any other type of unique identifier.
- LOCATION: includes references to the location of a person or center, without this representing a specific physical address, but rather generic locations relative to the name of a city/town, region or country.
- PERSON: includes names and surnames of people, as well as initials.
- REFERENCE: includes identifiers related to medical tests performed, such as clinical analyses, biopsies, scans or X-rays.

Table 2 shows the distribution of named entities from the Galén and MEDDOCAN corpora. For the different training, development/validation and test sets, the columns show the absolute number (abs) of entities named for each of the eight classes considered herein and their relative frequency (%). The majority NEs in the Galén corpus are related to Center, Person and Location, with approximately a percentage greater than 20% in all the considered sets. Meanwhile, the entities related to the Address are almost non-existent (2 in train, 1 in development/validation and 1 in test sets). With regard to the majority of the NEs in MEDDOCAN, they are associated with ADDRESS, PERSON and LOCATION, which also exceed a percentage of 20%. In this context, the most notable distinction between the corpora is the reduced prevalence of NEs pertaining to CENTER at MEDDOCAN, as opposed to ADDRESS, which constitutes the majority. It is also noteworthy that MEDDOCAN lacks NEs related to REFERENCE, such as identifiers of medical tests performed, which are exclusive to the Galén corpus.

Considering the collected entities, we needed to reorganize the entity classes identified by the PlanTL for the MEDDOCAN corpus. MEDDOCAN acknowledges 29 granular entity types, some of them with direct correspondence to Galén entity types. Table 3 shows the reorganization carried out indicating how the MEDDOCAN NEs were included within the Galén NEs. An asterisk (*) highlights instances where the same NE is included in two different types of Galén. Therefore, the entities labeled as 'TERRITORIO' (territory) that refer to a postcode were collected within the ADDRESS type, while the rest of the labeled entities are collected in LOCATION. Similarly, entities labeled as 'ID_SUJETO_ASISTENCIA' (id of patient) were associated with the HISTORY type when referring to health record numbers. Because MEDDOCAN documents have a semi-structured design, health record numbers are easily discernible due to their consistent presence with the accompanying text 'nhc'. Meanwhile, the other entities of 'ID_SUJETO_ASISTENCIA' were assigned to the entity type IDENT. The last entity that presents this particularity is FAMILIARES_SUJETO_ASISTENCIA (patient's family), which was associated to the PERSON type only when referring to an individual by name and excluded otherwise.

The granular entity types related in MEDDOCAN to dates, age, sex, profession and others were excluded from the study as they do not align with any entity types in Galén. Moreover, the entities corresponding to 'biometric', 'device' and 'vehicle' identifiers, 'internet protocol address' and 'url web', as well as 'beneficiary health plan number' have also been excluded due to their absence in the MEDDOCAN corpus. In other words, these entities do not appear in any text assigned to them

Table 3
Named entities considered by the PlanTL for the MEDDOCAN corpus and their assignment within the NEs considered for the Galén corpus. Asterisks refer to NEs whose association to each type is circumstantial based on their characteristics.

Galén NEs	MEDDOCAN NEs
ADDRESS	CALLE TERRITORIO*
CENTER	CENTRO_SALUD HOSPITAL INSTITUCION
CONTACT	CORREO_ELECTRONICO NUMERO_FAX NUMERO_TELEFONO
HISTORY	ID_SUJETO_ASISTENCIA*
IDENT	ID_SUJETO_ASISTENCIA* ID_TITULACION_PERSONAL_SANITARIO ID_EMPLEO_PERSONAL_SANITARIO ID_ASEGURAMIENTO
LOCATION	PAIS TERRITORIO*
PERSON	NOMBRE_SUJETO_ASISTENCIA NOMBRE_PERSONAL_SANITARIO FAMILIARES_SUJETO_ASISTENCIA*
Excluded	FECHAS EDAD_SUJETO_ASISTENCIA SEXO_SUJETO_ASISTENCIA PROFESION FAMILIARES_SUJETO_ASISTENCIA* ID_CONTACTO_ASISTENCIAL OTROS_SUJETO_ASISTENCIA
No presence	IDENTF_BIOMETRICOS IDENTIF_DISPOSITIVOS_NRSERIE IDENTIF_VEHICULOS_NRSERIE_PLACAS DIREC_PROT_INTERNET NUMERO_BENEF_PLAN_SALUD OTRO_NUMERO_IDENTIF URL_WEB

in the corpus. Table 2 shows the distribution after the re-assignation performed of NEs from the MEDDOCAN corpus.

2.2. Data augmentation

The nature of this task requires NLP models to recognize a large number of words assigned to entities that are not included in the vocabulary collected within the training corpus. When dealing with real-world problems, the corpus is typically small, increasing the probability of encountering words outside the vocabulary, which negatively affects the performance of the models. In order to mitigate this problem, a method for augmenting the availability of supervised documents is proposed in this study and described as follows.

For each document in the training dataset, a series of synthetic documents were generated by text surrogation of certain entities (those susceptible to natural replacement). Each synthetic document was generated by adding all the different paragraphs after the surrogation process. To avoid model over-fitting, if the same document was generated more than once, all the superfluous copies were purged out. The capitalization nature of the original caption was mimicked by the surrogated text. Entities which present numerical pattern (history numbers, references, phone or fax numbers and numerical identifiers) were replaced by perturbing digits randomly. Alpha-numeric entities were replaced by perturbing digits and characters except for certain parts (prefixes, protocol chains, etc.).

The surrogation of the entities consisted primarily of proper nouns (e.g. names, surnames, regions, countries) in a dictionary-based replacement process. Thus, a database was compiled from information supplied by the Spanish National Statistics Institute (INE).[1] Male and female names (52,287 available) and surnames (25,792 available) were picked randomly using a roulette wheel selection based on the absolute frequency of occurrence in the Spanish census. 99 regions (town, cities, provinces and autonomous regions) were picked randomly by using a uniform distribution. In order to simulate the probability of assisting to a medical center, countries (115 available) were picked by roulette wheel selection based on the absolute frequency of the registered immigrant population by country in the totality of the Spanish territory. Finally, medical center names were picked randomly from a directory of 871 public and private centers provided by the Spanish Ministry of Health.[2]

Table 4 illustrates the distribution of NEs in the Galén and MEDDOCAN augmented corpora when an augmentation denoted as x10 is applied, i.e., 10 surrogated copies were made for each document in the corpus. In addition, x20 and x30 augmentation levels were performed with for the two investigated corpora. However, the performance achieved with these augmentation were not significantly different from the x10 augmentation, these are therefore not included in the results section.

3. Methods

This section outlines the two distinct NER strategies proposed in this work to address the de-identification of real-world Spanish EHRs. Our first strategy employs RNN models, while our second one uses Transformer-based models to deal with the problem. In both cases, the de-identification task is approached as a sequence-labeling NER task, utilizing the IOB2 tagging scheme for token labeling [64].

3.1. Recurrent neural models

In this section, we outline the procedures involved in extracting features and generating embeddings from the input sequences, alongside the employment of RNNs to address the de-identification task.

3.1.1. Features

Given the relatively diminutive size of our training corpus, strategic decisions were made during the tokenization stage to curtail the vocabulary size while concurrently maximizing the extraction of contextual relationships. Notably, tokenization was conducted in a case-insensitive manner, and readily discernible expressions characterized by low error rates based on numerical sequences (e.g., DNI, NSS, CNP, dates, times, telephone numbers, references to analyses or samples, etc.) were systematically substituted with specialized tokens.

In recognition of the inherent significance of capitalization, especially in discerning acronyms and proper nouns, we introduced a series of descriptors to compensate for the lack of case distinction. These descriptors delineate aspects of the original word, explicitly marking the presence of initial capital letters, all-capital forms, lowercase forms, and digits.

Furthermore, we augmented tokens with flags signifying their attributes, denoting whether they represent a new line of text, a numeric expression, or function as separator characters. The tokenizer employed in this endeavor is a variant of NLTK's TokTok tokenizer.

To bolster our efforts, we leveraged a FastText skip-gram model as our embedding model, trained using the Galén corpus, which encompasses training documents and stored clinical records not earmarked for de-identification purposes. We optimized the FastText model parameters, specifying a word vector size of 400 and a window size of 10. This embedding methodology serves to mitigate the impact of

[1] https://www.ine.es/inebmenu/indiceAZ.htm
[2] https://www.sanidad.gob.es/ciudadanos/prestaciones/centrosServiciosSNS/

Table 4
Number and percentage of annotations per subset in Galén and MEDDOCAN corpora: training (T) and validation (V) with different DA percentages.

NE	Galén x10 DA				MEDDOCAN x10 DA			
	Train		Validation		Train		Validation	
	abs	%	abs	%	abs	%	abs	%
ADDRESS	22	.001	11	.003	18 711	.233	9394	.230
CENTER	5148	.315	1254	.296	3960	.049	2354	.058
CONTACT	451	.028	154	.036	5962	.074	2992	.073
HISTORY	198	.012	154	.036	6490	.081	3168	.078
IDENT	682	.042	99	.023	9515	.118	4609	.113
LOCATION	3762	.230	814	.192	19 206	.239	10 054	.246
PERSON	4828	.296	1617	.382	16 610	.206	8305	.203
REFERENCE	1235	.076	133	.031	0	.000	0	.000
Total	16 326		4236		80 454		40 876	

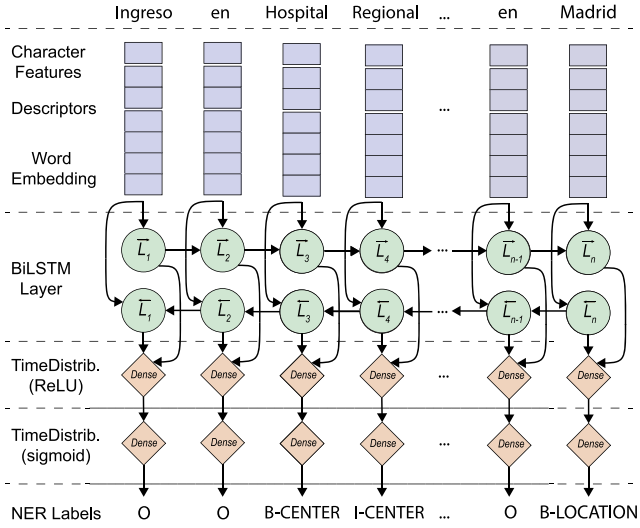


Fig. 2. The BiLSTM structure for the NER system comprises an embedding layer which includes descriptors, character level information and the word-embedding vector, that processes the input sequences. The LSTM forward layer is then computed based on the input and the previous state while the LSTM backward layer is computed based on the input and the future state (BiLSTM layer). The output of both layers is fed into a time-distributed layer utilizing ReLU and, ultimately, a time-distributed layer with sigmoid produces the tagging output.

multiple typographical errors that may be prevalent in a relatively confined collection of texts. Moreover, it facilitates the association of words expressed in a language with various inflectional nuances, characteristic of languages such as Spanish, which entail considerations of gender, number, and conjugation.

Consequently, we amassed three distinct categories of entry descriptors: character-level features (capitalization indicators), expression-level descriptors (new lines, separators, etc.), and word-embedding vectors. These descriptors are amalgamated within an embedding layer, forming the input dataset for the RNNs as illustrated in Fig. 2 (BiLSTM) and Fig. 3 (2-BiLSTM and BiLSTM-CRF).

3.1.2. Recurrent neural models

RNNs represent a generalization of feed-forward neural networks, tailored to tackle problems imbued with sequential structures. These networks are composed of hidden units and output units, intricately interconnected to form directed cycles. These connections give the units with an internal memory-like capacity, enabling the network to retain temporal context. In each iteration of the RNN training process, the current state of each hidden unit is estimated based on the input at the current time step and the previous hidden state, facilitating the processing of sequential information. The use of LSTM units [24] makes it possible to avoid the challenge of vanishing gradients in the network. LSTMs incorporate three distinct gates: input gates, responsible for determining which input values should be used to modify the memory block; forget gates, identifying features to be discarded; and output gates, governing the generation of the output.

Bidirectional LSTM (BiLSTM) comprises two LSTM networks, each trained to learn the representation of a token based on both its past and future context [28]. This bidirectional processing entails simultaneous examination of the sequence from left-to-right (past context) by one LSTM network and from right-to-left (future context) by another LSTM network. During each training step, a hidden unit forward layer ($\overrightarrow{L}$) is computed based on the current input and the preceding hidden state, while a hidden unit backward layer ($\overleftarrow{L}$) is estimated using the current input and the subsequent hidden state. The outcomes from both networks can be concatenated, yielding contextual information regarding the words surrounding each token. In our model, a Time Distributed layer at the output of the BiLSTM facilitates the return of the hidden state sequence to the subsequent layer, creating a k-node layer at each time step. The label for each input sequence is inferred by a final time-distributed layer that has a sigmoid activation function. Fig. 2 illustrates this RNN model. Additionally, the stacking of two Bidirectional LSTM layers (2-BiLSTM) enhances context awareness. In this configuration, the hidden unit forward layer of the second BiLSTM layer receives the output from the first layer, creating a more comprehensive contextual representation. Fig. 3(a) depicts the 2-BiLSTM model.

Another prevalent sequence model is the CRF, an undirected discriminative probabilistic graph model employed to represent probabilities pertaining to structured outputs based on input sequences. CRF model efficiently incorporates past and future entities to determine the probability of the output entity based on a set of input values. Its mechanism is similar to that of a bidirectional LSTM network employing past and future input features to mark sequences. The BiLSTM-CRF model combines the power of a BiLSTM and a CRF [22] for sequence learning and enhancing the performance of NER model. It takes a tokenized and embedded sequence as input, passing it through a BiLSTM layer. The result from this layer is then relayed to a time-distributed layer without an activation function. Additionally, to derive entity assignments for each word, a CRF model is incorporated into the output layer, as depicted in Fig. 3(b).

3.1.3. Transformers

In this study, the second strategy proposed to tackle the de-identification task leverages Transformer models. The foundational Transformer architecture, introduced by Vaswani et al. [29], harnesses self-attention to generate contextual numerical representations of each word, as well as to boost computational efficiency by parallelizing the network architecture. Over the last five years, Transformers have emerged as leading models in various areas of NLP [30,55], largely due to their efficacy in combination with transfer learning (TL) approaches. By following TL strategies, these models undergo pretraining on general-domain corpora and are then fine-tuned on specialized corpora for specific NLP tasks [30]. Employing Transformers with

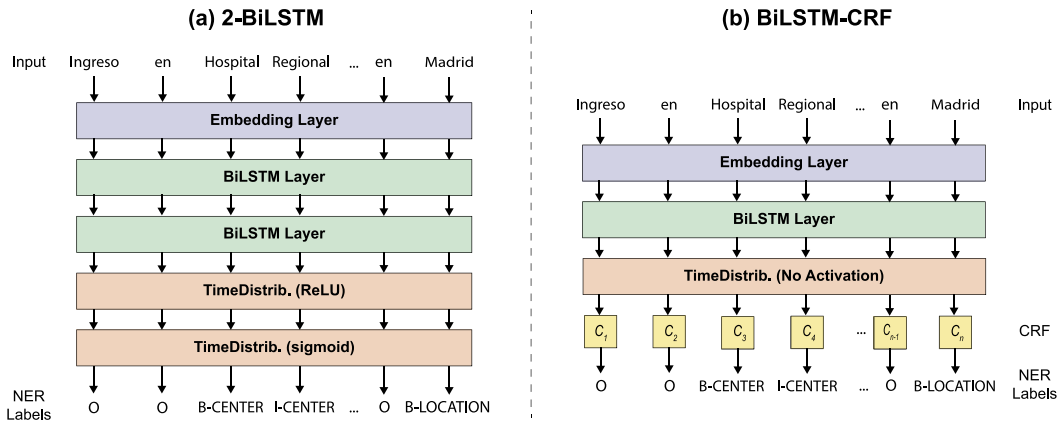


Fig. 3. The architectures of 2-BiLSTM model and BiLSTM-CRF model. (a) The 2-BiLSTM model. The LSTM forward layer of the second BiLSTM layer is connected to the output of the LSTM forward layer of the first BiLSTM layer, and the backward layer is connected to the backward layer of the first BiLSTM. (b) The BiLSTM-CRF model. The output of the time-distributed layer without activation function feeds the CRF model, which produces the tagging output.

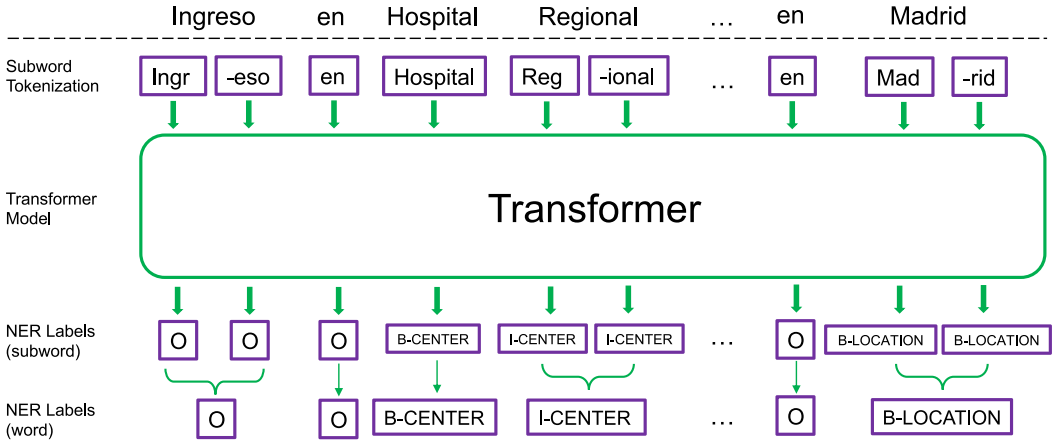


Fig. 4. Illustration of the Transformer-based methodology applied to tackle the de-identification problem.

diverse TL methodologies has led to SOTA performances in both the biomedical and clinical domains [31,34,65].

In this work, since we are dealing with the de-identification of medical texts in Spanish, we have used four distinct Transformer-based models that support the Spanish language, namely, XLM-RoBERTa (XLM-R), XLM-R-Galén, RoBERTa-BNE and RoBERTa-Bio:

- XLM-R: this multilingual version of the RoBERTa architecture [66] was pretrained on a massive general-domain 2.4TB CommonCrawl Corpus in 100 languages [33], using a large multilingual vocabulary of ~250K subwords. We experimented with both the Base (~277M trainable weights) and the Large (~559M trainable parameters) versions of the model.
- XLM-R-Galén: this model represents a domain-specific version of the XLM-R Base architecture. Specifically, it was obtained by performing a continual pretraining approach on a corpus of unlabeled real-world oncology clinical texts [34], with the aim of adapting the model to the particularities of the clinical domain in Spanish.
- RoBERTa-BNE: the general-domain Spanish version of the RoBERTa architecture [66] was pretrained on a 570 GB corpus obtained from the Spanish National Library (BNE) [36]. The model employs a Spanish vocabulary of ~50K subtokens and, again, we experimented with both the Base (~124M trainable parameters) and the Large (~354M trainable weights) versions of the model.
- RoBERTa-Bio: this Transformer-based model constitutes a biomedical pretrained language model for Spanish [35]. It was obtained by pretraining the RoBERTa Base architecture from scratch on several biomedical-clinical corpora in Spanish collected from publicly available resources, as well as a real-world clinical corpus collected from more than 278K clinical notes. The model uses domain-specific vocabulary in Spanish of 52K subwords.

An end-to-end methodology employing Transformers has been developed to tackle the problem of de-identifying real-world medical documents in Spanish. A visual depiction of the developed methodology is illustrated in Fig. 4. This methodology involves initially segmenting the text from medical documents into a sequence of subwords, since the Transformer-based models employed in this study process subword-level text as input data. Consequently, these subword sequences are fed into the models as the sole input, avoiding the need for additional input features. During the inference phase, given that Transformer-based models render predictions on a subword basis, it was necessary to map these predictions back to the word level, as depicted in Fig. 4. This was accomplished by adopting the maximum probability criterion outlined in [67], which involves assigning the label with the highest predicted probability to each word, based on the predicted probabilities obtained from the model on its constituent subwords. The resulting sequence of word-level labels could then be compared with the gold standard (GS) annotations to assess the performance of the models.

Since in this study we focus on a NER task, both RNN-based and Transformers-based methodologies have employed the IOB2 tagging scheme (see the output of the model in Fig. 2). Thus, for each NE considered in this work (see Section 2.1.2), two different labels were

generated: one with the "B-" prefix and the other with the "I-" prefix (e.g., for the "CENTER" category, the labels "B-CENTER" and "I-CENTER" were produced). Additionally, an "O" label was used for the words not belonging to any annotation. Consequently, since 8 NEs were considered, an output layer of 17 units — one unit for each possible label (i.e., $2 \times 8 + 1 = 17$) — was employed by all models applied in both methodologies.

4. Experiments and results

The results of the experiments conducted as well as the implementation details of the RNN-based and Transformer-based models used in this study are presented in this section. The experiments were repeated with 5 different seeds for all the analyzed models.

For the LSTM-based models, hyperparameter selection was performed by training the networks and using early-stop on the basis of the validation set. Thus, the final hyperparameters of these models were therefore chosen as those achieving the highest macro-averaged F1 score for the validation set. Finally, the performance evaluation of the different architectures was based on the predictions made by the RNN models on the test set. Thus, the resulting BiLSTM architecture proposed for this study is composed of 32 neurons in the bidirectional layer with 0.3 dropout, 16 neurons in the dense time-distributed layer with no dropout, and Rmsprop as the optimizer. The 2-BiLSTM network has a similar structure with 64 neurons in the first bi-layer and 0.3 dropout, 32 neurons in the second bi-layer and 0.2 dropout, and 16 neurons in the time-distributed layer. Finally, the BiLSTM-CRF model includes 128 neurons in the bi-layer with a 0.3 dropout, 8 neurons in the time-distributed layer (same as the entities considered) and Adam as the optimizer.

The standard method for evaluating the performance of a NER system is to compare the human annotations (ground truth) with the tagged output obtained. Depending on the comparison made, there are several ways to quantify the overall performance of the system. In this experiment, two are considered: the strict match and the exact match evaluations. For each NE, the type of entity identified and its boundaries are obtained, i.e. the character position where the NE begins to be labeled and the character position where it ends (i.e. the NE's span). In the strict match evaluation, both the type of entity and its spans must match, whereas in the exact match evaluation, only the span of the inferred entity is taken into account, not its type. The exact match evaluation is employed due to its relevance to the de-identification process. In a de-identification task, it is crucial that the concepts are identified entirely — i.e., their exact spans —, as this enables the removal of personal information from the document, even if the classification of the NE is not accurate.

We examine distinct groups of errors when evaluating our NER systems for de-identification. These groups were first presented at the MUC6 [14] and are founded on a comparison between the labeled NEs in the text (ground truth annotation) and the system output. The assessment considers four exclusive categories of error:

- Correct (cor): the NER system output and the ground truth annotation are the same.
- Incorrect (inc): the NER system output and the ground truth annotation do not match exactly.
- Missed (mis): the ground truth annotation was not captured by the NER system.
- Spurious (spu): the NER system infers an entity which does not correspond to a ground truth annotation.

The criteria for determining the accurate and inaccurate category differ based on the preferred performance evaluation of the NER system (strict/exact-match), as mentioned earlier. The metrics for evaluation — precision, recall and F1-score — are calculated assuming that the count of true positives (TP) is equal to the number of correctly identified entities. The count of NEs in the golden standard (gs_num) is equivalent to TP + false negatives, while the count of NEs in the NER system output (sys_num) is equivalent to TP + false positives.

Recall assesses the ability of a NER system to recognize all entities within a corpus (Eq. (1)), while precision evaluates the system's capacity to correctly identify entities (Eq. (1)). The F1-score is the harmonic mean of precision and recall (Eq. (2)), with the balanced F1-score being the most widely utilized.

$$\text{Precision} = \frac{\text{TP}}{\text{sys\_num}} \qquad \text{Recall} = \frac{\text{TP}}{\text{gs\_num}} \tag{1}$$

$$\text{F}_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

As many NER systems incorporate various entity types, it is often necessary to evaluate performance across all entity classes. Macro-averaged and micro-averaged F1-score are commonly used measures for this purpose. Macro-averaged F1-score calculates the F1-score independently for each entity type and then averages them (therefore treating all entity types equally). Micro-averaged F1-score aggregates the contributions of all entities from various classes to calculate the average (treating all entities equally). The latter measure is prevalent when evaluating NER models as it is a better indicator of model's performance in real application environments.

For reproducibility purposes of the experiments performed, the initial code needed to reproduce our research work on RNNs and Transformers is publicly available at https://github.com/guilopgar/DeIdentSpanishEHR.

4.1. De-identification results

Table 5 shows the outcomes of the strict evaluation — which represents the principal metric used to evaluate automatic de-identification systems [40,68] — obtained by the two strategies developed in this study to address the de-identification problem on the Galén real-world clinical corpus. Following the experimentation described above, the different RNNs and Transformers models were trained with each of the corpora — either separately or joined together (see '+' symbol in the tables) —, as well as with augmented versions ('x10' in the tables), and tested on the Galén corpus. Notice that the values (± 'between-validation performance' standard deviation) in Table 5 that are highlighted in bold indicate the best results achieved by all the models, while the underlined ones indicate the second best and those marked with the † symbol denote the best value attained by each independent model. The latter allows easy recognition of the training dataset with which the best results are achieved.

Table 5 shows that the best performance rate is obtained with the Transformers, namely the XLM-R (Large) model, that obtains values of 0.9415 ± *.01* precision, 0.9566 ± *.01* recall and 0.949 ± *.01* F1-score. In addition, it can observed that the models trained using only the MEDDOCAN corpus performed significantly worse. Thus, the highest F1-score value achieved is 0.6503 ± *.04* for XLM-R (Large) with x10 augmentation, which is below the lowest F1-score value achieved when training with the Galén corpus (without augmentation), which is 0.6824 ± *.02* with the BiLSTM model. Regardless of the training corpus, it can be observed in the table that increasing the corpus' size contributes to an improvement in the performance achieved by the de-identification NER models. Out of the 81 possible comparisons between metrics (augmentation vs. no augmentation), only one shows a higher performance for the non-augmented corpus, namely the precision obtained with MEDDOCAN-trained XLM-R (Base).

On the other hand, Table 6 shows the outcomes of the strict evaluation obtained by the RNNs and Transformers tested on the MEDDOCAN corpus. Again, the bold font in the table indicates the best results achieved, while the underlined ones indicate the second best and the ones indicated with the † symbol indicate the best value for each of the independent models. As with the Galén corpus, the XLM-R (Large) model, trained with the joined Galén+MEDDOCAN corpus and

Table 5
Micro-averaged metrics (mean ± sd from 5 independent experiments) computed on Galén's test set: precision (P), recall (R) and F1-score (F1) with strict evaluation. Models trained with different corpora (Galén, MEDDOCAN, Galén+MEDDOCAN and their augmented x10 versions, see leftmost column) and evaluated on Galen's test set. (See text for more details).

Train corpus	BiLSTM			2-BiLSTM			BiLSTM + CRF		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Galén	.6454 ± .02	.7240 ± .02	.6824 ± .02	.6598 ± .12	.7480 ± .06	.6995 ± .09	.8452 ± .02	.8011 ± .01	.8226 ± .01
x10	.7349 ± .02†	.8183 ± .01†	.7743 ± .01†	.7682 ± .04†	.8252 ± .02	.7955 ± .03	.8891 ± .02†	.8840 ± .02†	.8865 ± .02†
MEDDOCAN	.0208 ± .01	.1834 ± .07	.0371 ± .02	.0103 ± .01	.1377 ± .09	.0190 ± .02	.4550 ± .05	.3286 ± .04	.3809 ± .04
x10	.0558 ± .02	.2560 ± .06	.0911 ± .02	.0743 ± .02	.3623 ± .04	.1224 ± .02	.5748 ± .03	.4457 ± .03	.5016 ± .02
Gal + MED	.5765 ± .05	.7034 ± .02	.6333 ± .04	.6475 ± .04	.7349 ± .04	.6882 ± .03	.8220 ± .02	.7709 ± .02	.7956 ± .02
x10	.7122 ± .02	.8068 ± .02	.7566 ± .02	.7596 ± .04	.8405 ± .02†	.7979 ± .03†	.8638 ± .03	.8834 ± .01	.8734 ± .01
Train corpus	RoBERTa-Bio (Base)			RoBERTa-BNE (Base)			RoBERTa-BNE (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Galén	.8947 ± .01	.9080 ± .01	.9013 ± .01	.8630 ± .01	.8600 ± .01	.8615 ± .01	.8806 ± .02	.8914 ± .01	.8859 ± .01
x10	.9257 ± .01†	.9526 ± .01†	.9389 ± .01†	.8802 ± .01†	.9234 ± .01†	.9013 ± .01†	.9246 ± .01†	.9457 ± .01†	.9350 ± .01†
MEDDOCAN	.6219 ± .01	.4594 ± .02	.5283 ± .01	.4751 ± .03	.3440 ± .02	.3985 ± .02	.5965 ± .01	.3726 ± .01	.4586 ± .01
x10	.6481 ± .02	.4720 ± .03	.5459 ± .02	.5778 ± .08	.4326 ± .04	.4896 ± .03	.6348 ± .04	.4943 ± .02	.5553 ± .02
Gal + MED	.9070 ± .01	.9189 ± .01	.9129 ± .01	.8562 ± .01	.8469 ± .02	.8515 ± .01	.8995 ± .01	.8846 ± .02	.8919 ± .01
x10	.9243 ± .01	.9417 ± .01	.9329 ± .01	.8755 ± .02	.9120 ± .03	.8933 ± .02	.9205 ± .01	.9263 ± .02	.9234 ± .01
Train corpus	XLM-R-Galén (Base)			XLM-R (Base)			XLM-R (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Galén	.8703 ± .01	.8891 ± .01	.8796 ± .01	.8906 ± .01	.8846 ± .02	.8876 ± .02	.9209 ± .01	.9314 ± .01	.9261 ± .01
x10	.9138 ± .01†	.9326 ± .01†	.9231 ± .01†	.9136 ± .01	.9303 ± .01†	.9219 ± .01†	.9372 ± .01	.9554 ± .01	.9462 ± .01
MEDDOCAN	.5925 ± .02	.4440 ± .01	.5074 ± .01	.6465 ± .05	.4011 ± .02	.4943 ± .02	.6722 ± .02	.5217 ± .02	.5871 ± .01
x10	.6816 ± .02	.5240 ± .02	.5922 ± .02	.5776 ± .07	.5120 ± .05	.5402 ± .06	.7228 ± .03	.5937 ± .05	.6503 ± .04
Gal + MED	.8570 ± .01	.8760 ± .01	.8664 ± .01	.8858 ± .02	.8886 ± .02	.8871 ± .02	.9179 ± .01	.9269 ± .01	.9224 ± .01
x10	.9108 ± .01	.9229 ± .03	.9168 ± .02	.9166 ± .01†	.9234 ± .02	.9200 ± .01	**.9415 ± .01**†	**.9566 ± .01**†	**.9490 ± .01**†

Table 6
Micro-averaged metrics (mean ± sd from 5 independent experiments) computed on MEDDOCAN's test set: precision (P), recall (R) and F1-score (F1) with strict evaluation. Models trained with different corpora (Galén, MEDDOCAN, Galén+MEDDOCAN and their augmented x10 versions, see leftmost column) and evaluated on MEDDOCAN's test set. (See text for more details).

Train corpus	BiLSTM			2-BiLSTM			BiLSTM + CRF		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MEDDOCAN	.4667 ± .04	.6447 ± .03	.5398 ± .03	.6177 ± .06	.8327 ± .06	.7091 ± .06	.9101 ± .01	.8968 ± .00	.9034 ± .01
x10	.7509 ± .03	.7938 ± .04	.7716 ± .03	.7863 ± .02	.8500 ± .01	.8168 ± .01	.9150 ± .01	.9030 ± .01	.9089 ± .00
Galén	.2018 ± .06	.2189 ± .03	.2086 ± .04	.2341 ± .06	.2566 ± .05	.2426 ± .06	.3178 ± .02	.2423 ± .03	.2744 ± .02
x10	.2091 ± .06	.2482 ± .04	.2252 ± .04	.2371 ± .02	.2612 ± .06	.2475 ± .04	.3273 ± .04	.2643 ± .02	.2922 ± .03
Gal + MED	.4829 ± .10	.6236 ± .06	.5420 ± .09	.5890 ± .02	.7150 ± .03	.6455 ± .02	.9415 ± .01†	.9280 ± .00†	.9348 ± .00†
x10	.8159 ± .01†	.8542 ± .01†	.8346 ± .01†	.8503 ± .02†	.8761 ± .02†	.8630 ± .02†	.9267 ± .01	.9216 ± .00	.9241 ± .00
Train corpus	RoBERTa-Bio (Base)			RoBERTa-BNE (Base)			RoBERTa-BNE (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MEDDOCAN	.9723 ± .00	.9729 ± .00†	.9726 ± .00	.9561 ± .01	.9535 ± .00	.9548 ± .01	.9571 ± .00	.9572 ± .00	.9571 ± .00
x10	.9766 ± .00	.9717 ± .00	.9742 ± .00	.9648 ± .01†	.9636 ± .00	.9642 ± .00	.9710 ± .01	.9657 ± .00	.9683 ± .01
Galén	.4239 ± .02	.3911 ± .03	.4064 ± .02	.2743 ± .02	.2862 ± .02	.2801 ± .02	.3514 ± .03	.3022 ± .03	.3247 ± .02
x10	.4268 ± .03	.4114 ± .04	.4187 ± .04	.2747 ± .01	.3164 ± .02	.2940 ± .02	.3488 ± .05	.3675 ± .04	.3578 ± .04
Gal + MED	.9741 ± .01	.9717 ± .00	.9729 ± .00	.9647 ± .01	.9609 ± .00	.9628 ± .01	.9694 ± .00	.9642 ± .00	.9668 ± .00
x10	.9773 ± .00†	.9715 ± .01	.9744 ± .00†	.9707 ± .01	.9656 ± .00†	.9681 ± .00†	.9730 ± .00†	.9681 ± .01†	.9705 ± .00†
Train corpus	XLM-R-Galén (Base)			XLM-R (Base)			XLM-R (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MEDDOCAN	.9682 ± .00	.9652 ± .00	.9667 ± .00	.9763 ± .00	.9736 ± .01	.9749 ± .00	.9769 ± .00	.9744 ± .00†	.9756 ± .00
x10	.9741 ± .01	.9686 ± .00	.9713 ± .01	.9758 ± .00	.9714 ± .00	.9736 ± .01	.9777 ± .00	.9736 ± .00	.9756 ± .00
Galén	.3819 ± .03	.3884 ± .02	.3850 ± .03	.4005 ± .05	.3849 ± .03	.3923 ± .04	.4520 ± .04	.4400 ± .03	.4457 ± .04
x10	.3636 ± .01	.3960 ± .02	.3790 ± .01	.4623 ± .06	.4433 ± .05	.4523 ± .05	.4703 ± .05	.4731 ± .05	.4717 ± .05
Gal + MED	.9690 ± .00	.9659 ± .01	.9674 ± .00	.9752 ± .01	**.9749 ± .00**†	.9750 ± .00†	.9764 ± .00	.9742 ± .00	.9753 ± .00
x10	.9749 ± .01†	.9688 ± .00†	.9718 ± .01†	.9767 ± .00†	.9715 ± .01	.9741 ± .01	**.9782 ± .00**†	.9738 ± .00	**.9760 ± .00**†

x10 augmentation, achieves the best performance in terms of F1-score (0.9760 ± .00). This model also obtains the highest precision value (0.9782 ± .00). Nonetheless, the base version of the model trained with both corpora but without augmentation obtains the highest recall value (0.9749 ± .00) by less than 0.11%. It can also be observed that the performance achieved by the models trained with the Galén corpus is significantly worse. The XLM-R (Large) model obtains a maximum F1-score value of 0.4717 ± .04, a 6.8% less than the worst value obtained by training with MEDDOCAN. Finally, an increase in performance associated with the augmentation of the training data is noted. However, since MEDDOCAN is a medium-size corpus (1000 documents), the augmentation results in only a slight improvement for most models. Out of 81 metrics/models/training comparisons, there are 15 for which the non-augmented version performs better.

Additionally, Table 7 presents the outcomes of exact-match evaluation — spans of the inferred entity are correct, independently of the type of entity predicted — achieved by the best RNN and Transformer model in strict-match. The table shows the different metrics (precision, recall and F1-score) obtained by testing on both corpora and training the models with the x10 augmented corpus. In order to establish a baseline for comparison of the DL models, the results obtained on both corpora test sets with a rule-based system are also

Table 7
Micro-averaged metrics for strict and exact match evaluation strategies computed with a rule-based system and the best RNN and Transformer models when evaluated on the Galén and MEDDOCAN corpus. For each evaluation strategy, precision (P), recall (R) and F1-score (F1) metrics are computed. Highlighted are the best results achieved for each test corpus.

Model	Test corpus	Train corpus	NER (strict)			Spans (exact)		
			Precision	Recall	F1	Precision	Recall	F1
Rule-based	Galén	None	.4124 ± *.00*	.4171 ± *.00*	.4148 ± *.00*	.4153 ± *.00*	.4200 ± *.00*	.4176 ± *.00*
	MEDDOCAN	None	.6378 ± *.00*	.4491 ± *.00*	.5271 ± *.00*	.6416 ± *.00*	.4518 ± *.00*	.5302 ± *.00*
BiLSTM + CRF	Galén	Galén x10	.8891 ± *.02*	.8840 ± *.02*	.8865 ± *.02*	.8919 ± *.02*	.8868 ± *.02*	.8893 ± *.02*
		MEDDOCAN x10	.5748 ± *.03*	.4457 ± *.03*	.5016 ± *.02*	.6124 ± *.05*	.4748 ± *.03*	.5344 ± *.03*
		Gal + MED x10	.8638 ± *.03*	.8834 ± *.01*	.8734 ± *.01*	.8783 ± *.03*	.8983 ± *.01*	.8880 ± *.01*
	MEDDOCAN	Galén x10	.3273 ± *.04*	.2643 ± *.02*	.2922 ± *.03*	.4541 ± *.04*	.3671 ± *.03*	.4057 ± *.03*
		MEDDOCAN x10	.9150 ± *.01*	.9030 ± *.01*	.9089 ± *.00*	.9218 ± *.01*	.9097 ± *.01*	.9157 ± *.00*
		Gal + MED x10	.9267 ± *.01*	.9216 ± *.00*	.9241 ± *.00*	.9344 ± *.00*	.9292 ± *.00*	.9318 ± *.00*
XLM-R (Large)	Galén	Galén x10	.9372 ± *.01*	.9554 ± *.01*	.9462 ± *.01*	.9423 ± *.01*	**.9606 ± *.01***	.9513 ± *.01*
		MEDDOCAN x10	.7228 ± *.03*	.5937 ± *.05*	.6503 ± *.04*	.7794 ± *.04*	.6394 ± *.06*	.7008 ± *.04*
		Gal + MED x10	**.9415 ± *.01***	**.9566 ± *.01***	**.9490 ± *.01***	**.9454 ± *.01***	**.9606 ± *.01***	**.9529 ± *.01***
	MEDDOCAN	Galén x10	.4703 ± *.05*	.4731 ± *.05*	.4717 ± *.05*	.6389 ± *.04*	.6429 ± *.04*	.6408 ± *.04*
		MEDDOCAN x10	.9777 ± *.00*	.9736 ± *.00*	.9756 ± *.00*	.9789 ± *.00*	.9748 ± *.00*	.9769 ± *.00*
		Gal + MED x10	**.9782 ± *.00***	**.9738 ± *.00***	**.9760 ± *.00***	**.9794 ± *.00***	**.9751 ± *.00***	**.9772 ± *.00***

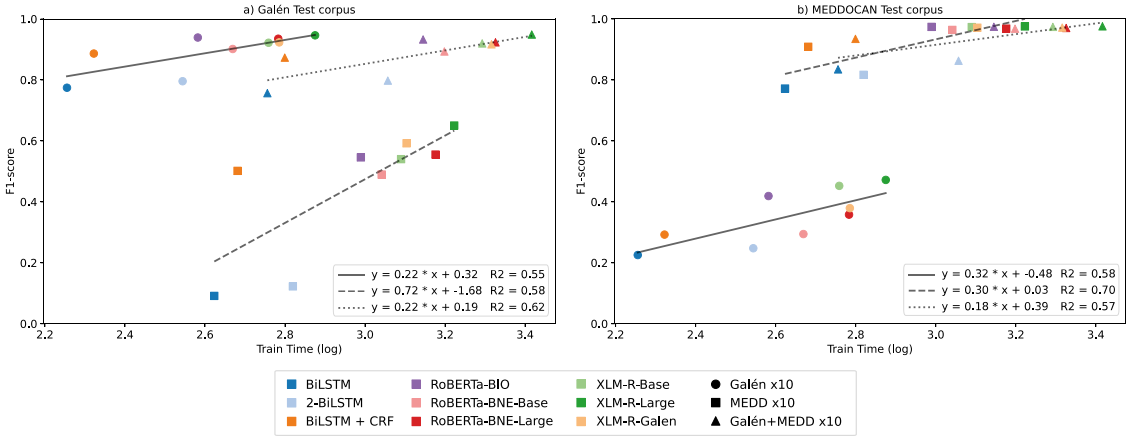


Fig. 5. Comparison of the F1-score performance achieved on Galén (a) and MEDDOCAN (b) test sets, with respect to the training time consumed by the models in a logarithmic scale. The NER strategies employed are represented with different colors, and the training set utilized is represented with different symbols. Linear regression is performed on the performance for each training set.

included. This system has been designed based on the rule-based model developed by López-Ubeda et al. [69] for the MEDDOCAN corpus. Additional rules have been included to adapt the system to the Galén corpus. The rule-based system does not require a training process, which has been indicated by the "None" label in the table.

As evident, the variance in F1-score performance achieved by the models between strict and exact-match evaluation is minimal in the cases of training with the same corpus or the joint corpus. Thus, we see unnecessary to exhibit the exact evaluation values for all conducted experiments. However, for the experiments with cross-corpus training (training with MEDDOCAN and test with Galén, and vice versa), the performance is increased by up to 17%. This suggests that the models are more proficient in determining the location of entities at text rather than their type. The XLM-R (Large) model once again yields the best outcome with an F1-score of 0.9772. It is noteworthy that both DL models markedly enhance the outcomes yielded by the rule-based system. This illustrates the efficacy of RNN and Transformers models in performing the NER task on unstructured documents, in contrast to rule-based systems.

Fig. 5 shows a comparison of the performance achieved by the various NLP strategies on Galén's (a) and MEDDOCAN's (b) test sets on the *y*-axis, with respect to the logarithmic computation time required for each of them to perform the training process on the *x*-axis. The various strategies are represented by different colors, while the symbols indicate the training set from which the F1-score value is obtained. The circles represent the performance achieved using Galén's train set with DA, the squares represent MEDDOCAN's train set with DA, and the triangles represent the union of the two previous sets. Furthermore, Fig. 5 illustrates a linear regression on the F1-score achieved for each training data set, thereby enabling an examination of the direct impact of the training time on the performance of the NER strategy. The legend shows the function and the coefficient of determination (R^2). The solid line represents the results with Galén, the dashed line MEDDOCAN, and the dotted line the join of both corpus.

The RNN bi-lstm requires the least computational time, with 180 min (3 h) of training with Galen's set. In contrast, the Transformer XLM-R-Large requires the longest time, with 2100 min (35 h) of training with Galen+MEDDOCAN's set. All experiments have been executed on a PC running under Linux Ubuntu 22.04.5 LTS, equipped with an AMD Ryzen ThreadRipper Pro 3955WX 3.9 GHz, 128 GB RAM, NVIDIA® Geforce RTX 3090 GPU with 24 GB RAM.

Fig. 5 shows the clusters of F1-score values according to the training set utilized, with the exception of the MEDDOCAN's test set case (b), wherein the values obtained with the MEDDOCAN x10 sets and the union of Galen + MEDDOCAN exhibit some overlap. This suggests that there is a reduced discrepancy in performance and running time with both sets. In terms of the required training times, the RNNs require less computational time than the Transformer models, as expected due to their lower complexity. Among the Transformers, the shortest training times are observed for those based on RoBERTa (RoBERTa-BIO, RoBERTa-BNE-Base/Large) in comparison to their equivalent models based on XLM-R (XLM-R-Galén, XLM-R-Base/Large). This is directly

Table 8
NER (strict-match) metrics for each NE obtained by the recurrent network BiLSTM + CRF system tested on the Galén's test set when trained using the joined Galén+MEDDOCAN x10 augmented corpus. Number of NEs inferred correctly (cor) and incorrectly (inc), missed (mis) or spurious (spu), and the total number of NEs in the ground-truth (GT) set and inferred by the de-identification model (infer).

NE	Precision	Recall	F1	COR	INC	MIS	SPU	GT	INFER
ADDRESS	1.000	1.000	1.000	1	0	0	0	1	1
CENTER	.8842	.9231	.9032	84	9	0	3	91	95
CONTACT	.9412	.9412	.9412	16	1	0	0	17	17
HISTORY	.9167	.7333	.8148	11	4	0	1	15	12
IDENT	.5833	.8750	.7000	7	1	0	4	8	12
LOCATION	.8448	.7424	.7903	48	10	9	6	66	58
PERSON	.8561	.9154	.8848	119	12	2	2	130	139
REFERENCE	.8800	1.000	.9362	22	0	0	3	22	25
macro-avg	.8611	.8894	.8693						
micro-avg	.8579	.8800	.8688	314	31	11	19	350	359

Table 9
NER (strict-match) metrics for each NE obtained by the XLM-R (large) system tested on the Galén's test set when trained using the joined Galén+MEDDOCAN x10 augmented corpus. Number of NEs inferred correctly (cor) and incorrectly (inc), missed (mis) or spurious (spu), and the total number of NEs in the ground-truth (GT) set and inferred by the de-identification model (infer).

NE	Precision	Recall	F1	COR	INC	MIS	SPU	GT	INFER
ADDRESS	1.000	1.000	1.000	1	0	0	0	1	1
CENTER	.9579	1.000	.9785	91	0	0	4	91	95
CONTACT	1.000	1.000	1.000	17	0	0	0	17	17
HISTORY	.8824	1.000	.9375	15	0	0	2	15	17
IDENT	1.000	1.000	1.000	8	0	0	0	8	8
LOCATION	.9531	.9242	.9385	61	3	2	1	66	65
PERSON	.9542	.9615	.9579	125	4	1	2	130	131
REFERENCE	.9565	1.000	.9778	22	0	0	1	22	23
macro-avg	.9630	.9857	.9738						
micro-avg	.9551	.9714	.9632	340	7	3	10	350	357

related to the number of parameters of each model, with RoBERTa-BNE-Base and Large having ∼124M and ∼354M trainable parameters, respectively, and XLM-R-Base and Large having ∼277M and ∼559M trainable parameters, respectively. Consequently, RoBERTa-BNE-Large required a training time between that of the XLM-R-Base and XLM-R-Large models.

4.2. Metrics for each NE

In order to conduct a comprehensive evaluation of the performance of BiLSTM + CRF RNN system and XLM-R (Large) Transformer — which achieved the highest de-identification results (see Table 7) — Tables 8 and 9 present the outcomes with these NER model for each NE separately. Apart from the F1-score, following the "risk aversion" principle, recall is often considered the reference metric to evaluate the performance of de-identification systems in clinical settings, as false negative errors may have an impact on the privacy of patients and medical professionals [68]. An automatic system demonstrating a recall score above 0.95 is generally regarded as reliable in its ability to de-identify a clinical corpus [68,70].

As we can see from Table 9, the XLM-R Transformer achieves a recall value over 0.95 in 7 of the 8 NEs, proving the viability of the model as a reliable medical de-identification system. The NE for which the model does not reach a recall value of 0.95 is LOCATION (with 0.9242 recall). In general, the model demonstrates excellent recall in the strict-match evaluation as it only missed 3 out of the total 350 entities in the ground truth set. Additionally, only 7 entities were incorrectly predicted due to wrong span or entity type given by the model. Finally, 10 spurious entities were inferred by the model.

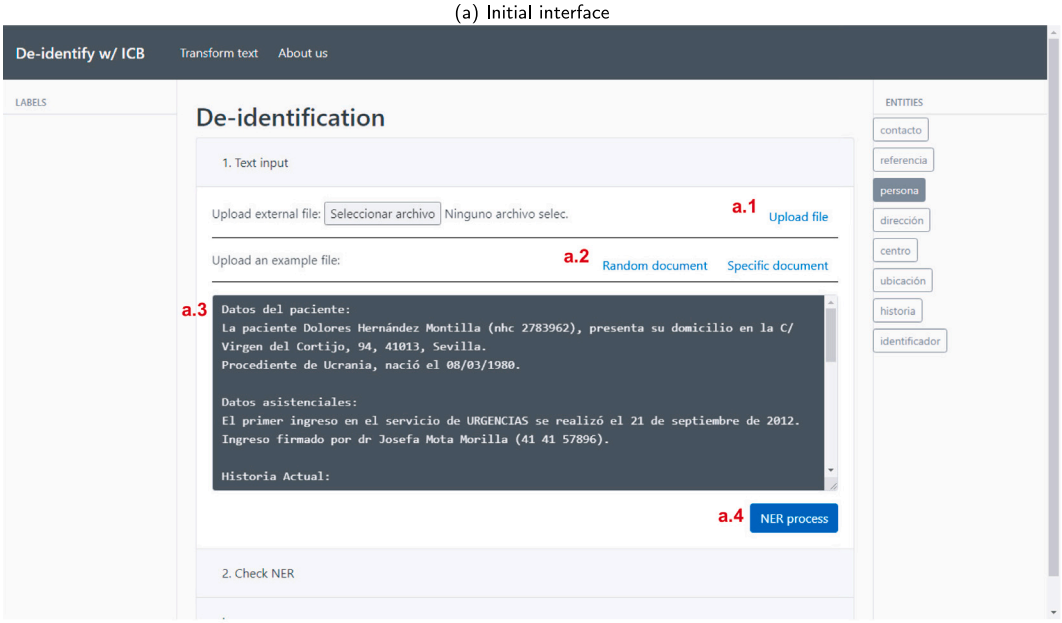
In contrast, Table 8 shows the results obtained by the BiLSTM + CRF system. This model exhibits a recall value higher than 0.95 for only 2 of the 8 NEs, being lower than 0.90 for 3 NEs. An examination of the evaluation metrics reveals that the model exhibits a markedly lower recall since it missed 11 out of the 350 entities. Additionally, the model inferred 19 spurious entities, nearly twice the number inferred by the Transformer model. However, its suboptimal performance is evidenced by the number of entities that were incorrectly classified. A detailed examination of these errors reveals that the majority of them are the result of confusions between LOCATION, CENTER, and PERSON entities. The RNN demonstrates a lesser capacity for generalization and the acquisition of contextual information than Transformers. There are words that alone are insufficient for clearly identifying the type of NE in question. To illustrate this, the term "Dolores" in Spanish may represent a person's first name, or be included in the name of a medical facility, or refer to the patient's sensation of pain. Furthermore, the model demonstrates a similar bias in the classification between IDENT and HISTORY, which is not the case with Transformers.

5. De-identify application

Considering the outstanding performance accomplished by some of the NER models analyzed in this work, an application was designed to use these models for the automatic de-identification of clinical documents written in Spanish. The software, built with HTML, Javascript and Flask, establishes a platform for uploading and editing files. These files can be processed using the NLP model and anonymized in compliance with data protection laws. It is worth noting that the tool has been configured for use with any NER model, allowing the inference module to be changed according to factors such as computational power available. Transformer-based models demand high levels of computational performance when compared to neural networks, which is not easily achievable at health information setups. The figures illustrating the performance of the tool were acquired using the XLM-R (Large) model, which has been trained with the Galén+MEDDOCAN x10 joint corpus. Fig. 6 shows the application's interface, depicting the initial state (6.a) and the state after applying the de-identification model (b). The tool allows for the initial uploading of txt, MS Word®, or pdf files to be processed (6.a.1) or for the selection of a document from the database (6.a.2). Once the document has been selected, its content can be easily edited in panel 6.a.3. When the text content is as desired, the NER process can then be initiated by pressing button 6.a.4.

Once the NER process is complete, the interface changes to the one depicted in Fig. 6(b), where three panels are visible. Panel 6.b.3 on the right is shared with the initially shown interface in Fig. 6(a). In the

(a) Initial interface



(b) NER process interface

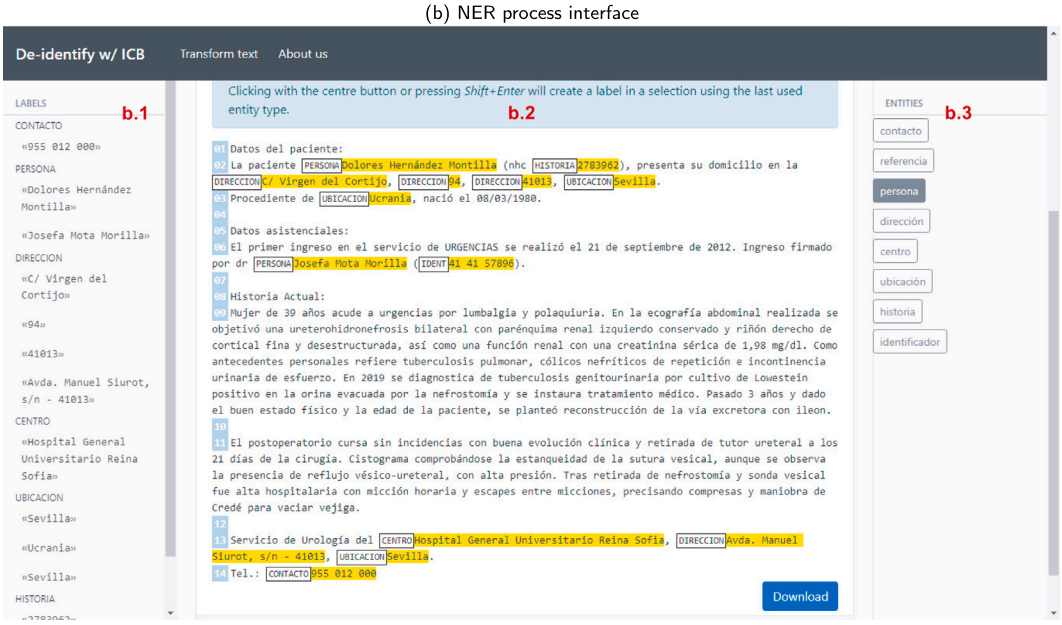


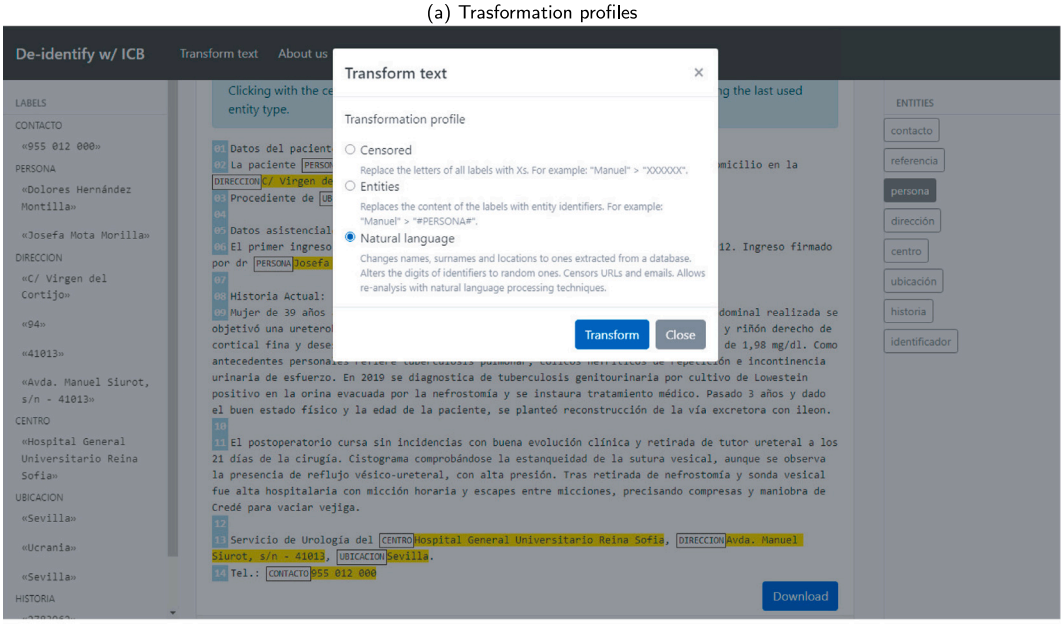
Fig. 6. Interface of the application developed to perform de-identification of documents, (a) initially and after a document has been uploaded to the application, and (b) after the NER process has been performed on the selected text.

central panel (6.b.2) we can see a version of the web-based annotation tool 'brat'. The panel displays the content of the previously selected and edited text, showing the NEs recognized by the NER model. The entity type appears in boxes beside the associated text section, which is underlined. All entities that have been identified are displayed in the left panel (6.b.1), organized by their type and providing quick access to them through a single click. The central panel enables users to modify the labels by removing or altering them when identifying potential errors in the NER process. As an example, user can replace a 'LOCATION' (*ubicación* in Spanish) label with a 'CENTER' (*centro* in Spanish) label. Furthermore, the panel supports creating new labels for specific sections of text by selecting the desired word(s) with the mouse, then clicking on the entity in the right panel (6.b.3) for assignment. This function enables users to easily monitor and correct any errors that the de-identification model may have made during the NER process. When the text displays the desired labels, the tool allows the content to be retrieved by clicking on the corresponding 'save' button. The software downloads a compressed file to the user's computer which contains a text document of the processed content and an annotations file (.ann format) highlighting identified entities, relevant text sections and their corresponding start/end positions in the main document.

The tool provides an additional feature, the transformation of text for its anonymization. By clicking the link in the navbar ('*Transform text*'), a modal opens in which the user can choose one of three available transformation profiles, as showed in Fig. 7(a):

- Censored: this method replaces the letters of all words composing the labeled text section with 'X'. In this way, the text retains the information about the original position of all the words, but not the content of the labeled sections.

(a) Trasformation profiles



(b) De-identified text

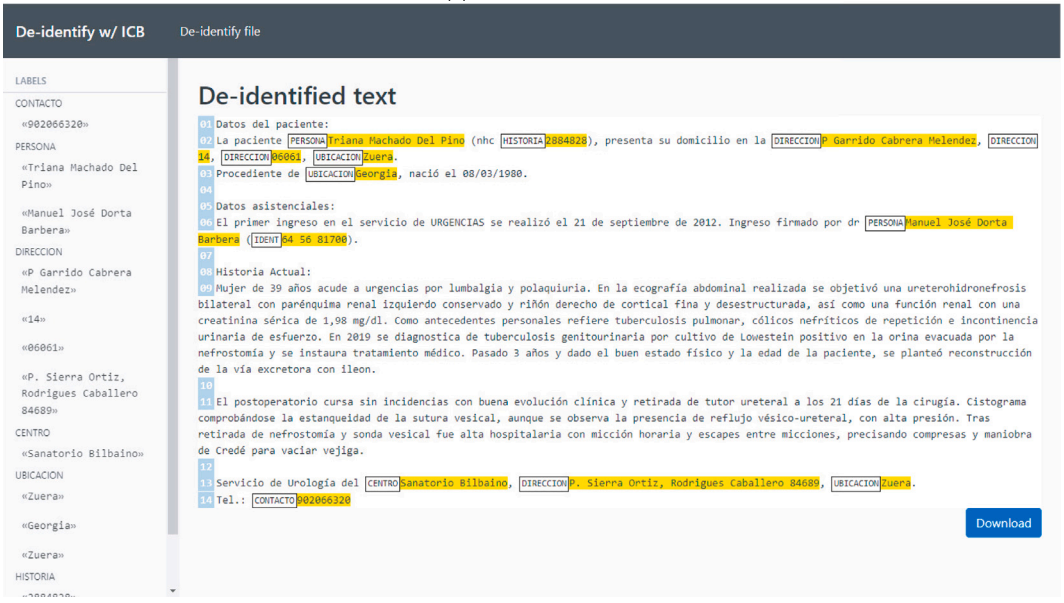


Fig. 7. Interface of the application shown (a) when the transformation process (de-identification) of the text previously processed by the NER model is started, and (b) the text with the entities already transformed with natural language profile.

- Entities: this method replaces the text with the content of the entity, i.e. a name in the text is replaced by '#PERSON#'. This profile allows the general context of the sentences to be maintained.

- Natural language: This method replaces words associated with a label with other words that respond to the same context of the label. Basically, this transformation profile replaces the names, surnames, medical centers and locations with other information sourced from a database compiled from data supplied by the Spanish National Statistics Institute (INE)[3] and the Spanish Ministry of Health.[4] Male and female names, surnames and countries were picked randomly using a roulette wheel selection based on the absolute frequency of occurrence in the Spanish census. Regions and medical center names were picked randomly by using a uniform distribution. In this way, the aim is to maintain consistency in the de-identified set of documents. Additionally, the software aims to maintain a parallel structure, such that a three-word name (consisting of a first name and two surnames) is transformed into another three-word name with replacements for each one. Entities containing numbers, such as identifiers, history numbers, or references, are randomly altered. Addresses, on the other hand, exhibit a mixture of both behaviors, replacing words that indicate the name of streets and avenues, and altering digits of builds and stairs numbers. This natural language profile facilitates the use of other NLP techniques on de-identified text.

[3] https://www.ine.es/inebmenu/indiceAZ.htm
[4] https://www.sanidad.gob.es/ciudadanos/prestaciones/centrosServiciosSNS/

Fig. 7(b) illustrates the interface that emerges after selecting and performing the intended transformation. Similar to Fig. 6(b), the left panel indicates every identified entity in the text that has been transformed, while the center panel shows the corresponding transformation outcome. It is worth noting that unlike the previous interface, no text editing and entity modifications are feasible. The transformed text can be downloaded in a txt file format, allowing the user to edit the text and labels again or perform a new transformation by returning to the interface shown in Fig. 6(b). For illustration purposes, the figures show an example of a real-world medical document in Spanish. This document was elaborated with simulated information, thus complying with the Spanish LOPD-GDD and the European GDPR. As we can see in Fig. 6(b), for this particular clinical document, the XLM-R (Large) Transformer-based model was able to correctly identify all PHI entities contained in it. This tool is public (on demand by a user) on the url https://www.icb.lcc.uma.es:33898.

Although the software is currently distributed to medical managers at the *Regional* and *Virgen de la Victoria* University Hospitals in Málaga (Spain) for validation, its future integration into the Galén [39] information system will enable it to be interconnected with a real-world medical database. The application will allow direct patient import from the database to modify medical appointments in real time, preserving the Spanish LOPD-GDD and the European GDPR, and then saving de-identified copies for clinical studies. Moreover, an automatic de-identification mode for batches of documents will be included to remove private, sensitive information from the EHRs stored on previous years. It is of note that the integration of the application within the health system will facilitate the control of the consistency of the de-identified reports. The system will record the actions performed by each user in a private manner. Such actions will include a data linkage between the transformed data and the original data via identifiers. This will allow clinical staff to validate and verify the process performed, with the option of returning the documents to their original state. Additionally, by making slight adjustments to the previous method and obtaining labeled texts instead of de-identified ones, a cyclic process can be maintained for the improvement of the developed de-identification models. Under expert supervision, more labeled documents could be gathered through a semi-automatic approach, thereby expanding the corpus size. Obtaining quality data is the biggest hurdle in this type of research and this tool can help obtain this data with significantly less effort.

6. Discussion

The results described in Tables 5 and 6 lead to various conclusions. First, the expected higher performance on the NER task of the Transformers models compared to the RNNs models. Testing on the Galén corpus, the highest F1-score value achieved by the Transformer-based model XLM-R (Large) is 0.949, 7.6% better than the highest value achieved by an RNN model, i.e. 0.8734 with BiLSTM + CRF. Similarly, testing on the MEDDOCAN corpus, the highest F1-score value reached by that same Transformer is 0.976, 4.1% better than the 0.9348 value achieved with BiLSTM + CRF. This is supported by the results in exact evaluation (only spans) shown in Table 7. The results presented in Table 7 also demonstrate the markedly superior performance of these models in comparison to rule-based systems when processing unstructured text.

When considering the Transformers, it is noteworthy that while the RoBERTa-BNE model was pretrained on an exclusive corpus of Spanish texts, the XLM-R model outperforms the RoBERTa-BNE in the task of de-identifying Spanish medical cases. Therefore, for this specific task, the large-scale multilingual pretraining carried out by the XLM-R model has yielded superior results than the Spanish-specific pretraining employed by the RoBERTa-BNE model. Moreover, we conclude that the large versions of the Transformers XLM-R and RoBERTa-BNE models outperform their base models. RoBERTa-Bio is the only base model with a comparable performance, achieving an F1-score of 0.9329 on Galén and 0.9744 on MEDDOCAN corpora, in all cases having been trained with the Galén+MEDDOCAN x10 joint corpus. This can be explained by the fact that RoBERTa-Bio is a biomedical-domain model that was trained using Spanish biomedical corpora of 1.1 billion tokens and an EHR corpus of 95 million tokens. Regarding the RNN models, the BiLSTM + CRF network is the top-performing RNN model for both test corpora, outperforming 2-BiLSTM by 6.7%. The difference compared to the previous research [38], where the 2-BiLSTM outperformed, can be attributable to the in-depth and extensive experimentation conducted in this study.

Regarding the performance achieved by the models when comparing the induced-efficiency supplied by the corpora used to train them, according to the F1-score all NER models applied in this work benefit from the DA procedure. Only in 3 out of 54 experiments, a better performance was obtained with a non-augmented corpus. This reinforces the fact that increasing variability and vocabulary can enhance the performance of NER models. This is also inferred from the superior performance results with models trained with the joined (Galén+MEDDOCAN) corpora, rather than the independent corpora. Moreover, it is observed that cross-training the corpora does not give sufficient performance. The models are incapable of correctly recognizing entities in the texts due to the particularities of both corpora. Documents collected at MEDDOCAN are clinical cases of medical articles with semi-structured information, while the documents from Galén consist of unstructured texts acquired from real-world medical activities. The results in the Table 7 summarize and exemplify these conclusions.

As shown in the Table 9, the majority of the mistakes were made by the model at LOCATION, PERSON and CENTER entities. Regarding the mistakes in identifying LOCATION and CENTER entity types, these are attributed to the model's difficulty in distinguishing between them. The Spanish Ministry of Health dictates that a medical center's name must include its location as part of its name whenever the name of the center itself is not sufficient to identify it unequivocally. For instance, in the text "Hospital Universitario Regional de Málaga", the name of the Spanish city "Málaga" is labeled as part of the name of the center due to the abundance of medical centers named "Hospital Universitario Regional" in the Spanish registry. However, in the text "Hospital Universitario Virgen de la Victoria de Málaga", "Málaga" is labeled as a location rather than being part of the name of the medical center, and "de" is labeled as O type (in the IOB2 tagging scheme). This poses a challenge not only for automated models but also for human annotators.

The efficiency in terms of computational time required for training the RNNs and Transformers models with the various corpora has also been evaluated. Fig. 5 illustrates that, in general, an increase in computational time is required to achieve an improvement in F1-score performance. The linear regression coefficients are all positive, exceeding 0.22 in all but one case, thereby indicating a gain for all the training sets under analysis. Two models stand out as being significantly more effective than the rest. The first of these is the BiLSTM + CRF, which exhibits a higher level of performance than the 2-BiLSTM model, but requiring less computational time. The second is RoBERTa-BIO, which also requires less computational time than the others Transformers, obtaining a significant F1-score that is close to the best. In conclusion, it is evident that although Transformer-based models with a greater number of trainable parameters demonstrate superior performance, this is accompanied by a significant increase in computational time during training. It should be noted that there is minimal difference in inference times due to the implicit parallelism of Transformers. However, they do require greater hardware resources, particularly in terms of RAM.

We have performed additional experiments to thoroughly analyze the performance of the Transformer models in the cross-corpus setting. The results are described in Tables 10 and 11. Thus, we have analyzed the performance of the systems when trained on both the Galén and

Table 10
Micro-averaged metrics computed on Galén's test set. We report the performance of each Transformer model when trained on both the Galén and the Galén+MEDDOCAN corpora, as well as when following the TL approach, i.e., the model is firstly pretrained on the MEDDOCAN corpus and then fine-tuned on the Galén corpus. "REFERENCE" label was eliminated from the Galén corpus, since it is not contained in the MEDDOCAN corpus (see Table 2). Finally, for strict evaluation strategy, precision (P), recall (R) and F1-score (F1) metrics are computed.

	RoBERTa-Bio (Base)			RoBERTa-BNE (Base)			RoBERTa-BNE (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Galén	.8938 ± .01	.9030 ± .01	.8984 ± .01	.8682 ± .01	.8549 ± .02	.8614 ± .01	.8802 ± .01	.8860 ± .01	.8831 ± .01
x10	.9269 ± .01	.9494 ± .01	.9380 ± .01	.8831 ± .01	.9213 ± .01†	.9018 ± .01†	.9257 ± .01	.9421 ± .01†	.9338 ± .01
Gal + MED	.9085 ± .01	.9134 ± .01	.9109 ± .01	.8606 ± .01	.8427 ± .02	.8515 ± .02	.8973 ± .01	.8854 ± .02	.8913 ± .01
x10	.9249 ± .03	.9378 ± .02	.9313 ± .02	.8781 ± .02	.9091 ± .03	.8933 ± .02	.9225 ± .01	.9213 ± .02	.9219 ± .02
MED → Gal	.9031 ± .01	.9091 ± .01	.9061 ± .01	.8716 ± .01	.8610 ± .02	.8662 ± .01	.9182 ± .01	.8896 ± .01	.9037 ± .01
x10	.9302 ± .02†	.9500 ± .01†	.9400 ± .01†	.8873 ± .02†	.9159 ± .02	.9013 ± .02	.9345 ± .01†	.9396 ± .02	.9371 ± .01†
	XLM-R-Galén (Base)			XLM-R (Base)			XLM-R (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Galén	.8738 ± .01	.8866 ± .01	.8801 ± .01	.8975 ± .02	.8768 ± .01	.8870 ± .02	.9213 ± .01	.9280 ± .01	.9247 ± .01
x10	.9147 ± .01	.9287 ± .01	.9216 ± .01	.9122 ± .01	.9256 ± .01†	.9189 ± .02†	.9359 ± .01	.9524 ± .01	.9441 ± .01
Gal + MED	.8673 ± .02	.8762 ± .01	.8717 ± .01	.8847 ± .02	.8811 ± .01	.8828 ± .02	.9186 ± .01	.9226 ± .02	.9206 ± .02
x10	.9099 ± .01	.9177 ± .02	.9138 ± .01	.9149 ± .02	.9183 ± .01	.9166 ± .02	.9422 ± .01	.9537 ± .01	.9479 ± .01
MED → Gal	.8753 ± .01	.8817 ± .01	.8785 ± .01	.8971 ± .01	.8695 ± .01	.8830 ± .01	.9257 ± .01	.9268 ± .02	.9263 ± .01
x10	.9266 ± .01†	.9470 ± .02†	.9367 ± .01†	.9172 ± .02†	.9183 ± .01	.9177 ± .01	**.9469 ± .01**†	**.9573 ± .01**†	**.9521 ± .01**†

Table 11
Micro-averaged metrics computed on MEDDOCAN's test set. We report the performance of each Transformer model when trained on both the MEDDOCAN and the Galén+MEDDOCAN corpora, as well as when following the TL approach, i.e., the model is firstly pretrained on the Galén corpus and then fine-tuned on the MEDDOCAN corpus. "REFERENCE" label was eliminated from the Galén corpus, since it is not contained in the MEDDOCAN corpus (see Table 2). Finally, for strict evaluation strategy, precision (P), recall (R) and F1-score (F1) metrics are computed.

	RoBERTa-Bio (Base)			RoBERTa-BNE (Base)			RoBERTa-BNE (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MEDDOCAN	.9723 ± .00	.9729 ± .01	.9726 ± .00	.9561 ± .01	.9535 ± .00	.9548 ± .01	.9571 ± .01	.9572 ± .01	.9571 ± .01
x10	.9766 ± .01	.9717 ± .00	.9742 ± .01	.9648 ± .01	.9636 ± .01	.9642 ± .01	.9710 ± .00	.9657 ± .01	.9683 ± .00
Gal + MED	.9743 ± .00	.9717 ± .00	.9730 ± .00	.9647 ± .00	.9609 ± .01	.9628 ± .00	.9695 ± .01	.9642 ± .00	.9669 ± .00
x10	.9775 ± .01†	.9715 ± .01	.9745 ± .01	.9708 ± .00†	.9656 ± .01	.9682 ± .01	.9731 ± .01†	.9681 ± .00†	.9706 ± .01†
Gal → MED	.9750 ± .01	.9747 ± .01	.9749 ± .01	.9614 ± .01	.9594 ± .00	.9604 ± .01	.9720 ± .01	.9680 ± .01	.9700 ± .01
x10	.9771 ± .00	.9747 ± .01†	.9759 ± .01†	.9702 ± .00	.9664 ± .01†	.9683 ± .01†	.9715 ± .01	.9670 ± .00	.9692 ± .01
	XLM-R-Galén (Base)			XLM-R (Base)			XLM-R (Large)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MEDDOCAN	.9682 ± .01	.9652 ± .01	.9667 ± .01	.9763 ± .00	.9736 ± .01	.9749 ± .01	.9769 ± .01	.9744 ± .01	.9756 ± .01
x10	.9741 ± .01	.9686 ± .01	.9713 ± .01	.9758 ± .00	.9714 ± .01	.9736 ± .00	.9777 ± .01	.9736 ± .00	.9756 ± .01
Gal + MED	.9691 ± .00	.9659 ± .01	.9675 ± .01	.9753 ± .00	.9749 ± .01	.9751 ± .01	.9766 ± .01	.9742 ± .00	.9754 ± .01
x10	.9749 ± .01†	.9688 ± .00	.9718 ± .01	.9767 ± .00	.9715 ± .01	.9741 ± .01	.9783 ± .01	.9738 ± .01	.9761 ± .01
Gal → MED	.9697 ± .01	.9682 ± .01	.9689 ± .01	.9782 ± .01	.9749 ± .01	.9765 ± .01	**.9794 ± .01**†	.9757 ± .00	.9775 ± .01
x10	.9746 ± .01	.9695 ± .00†	.9721 ± .01†	.9788 ± .01†	.9749 ± .01†	.9768 ± .01†	.9792 ± .01	**.9759 ± .01**†	**.9776 ± .01**†

MEDDOCAN corpora and evaluated on the Galén corpus (see Table 10), and also when trained on both corpora and evaluated on the MEDDOCAN test set (see Table 11). Additionally, we have also analyzed the performance of the Transformer models following an alternative TL approach, in which the models were pretrained on a *base* dataset and then fine-tuned on a *target* dataset. Thus, in Table 10, the *base* dataset corresponds to the MEDDOCAN corpus, while the *target* dataset is Galén, given that, in this case, Galén is the corpus used to evaluate the performance of the models. Conversely, in Table 11, the Galén corpus corresponds to the *base* set, while MEDDOCAN is the *target* set. For all experiments conducted to produce the results shown in Tables 10 and 11, in order to maximize the compatibility between the Galén and the MEDDOCAN corpora, the "REFERENCE" label was removed from the Galén corpus, since it is not contained in the MEDDOCAN corpus (see Table 2).

Table 10 shows that the best performance in the Galén corpus is achieved by the XLM-R (Large) model when following the TL approach using the augmented versions of both corpora, obtaining micro-averaged values of 0.9469, 0.9573 and 0.9521 in precision, recall and F1-score, respectively. In 4 of the 6 analyzed Transformer models, the best micro-averaged F1-score is obtained when following the TL approach using the augmented corpora. In Table 11, the same pattern is observed. Hence, the best performance in the MEDDOCAN corpus is also obtained by the XLM-R (Large) model when following the TL approach using the augmented versions of both corpora, obtaining micro-averaged values of 0.9792, 0.9759 and 0.9776 in precision, recall and F1-score, respectively. In 5 of the 6 Transformer models examined herein, the highest micro-averaged F1-score is obtained when following the TL strategy using the augmented corpora.

Given the results obtained in both Tables 10 and 11, we can conclude that, in the cross-corpus setting, the TL-based approach led the Transformer models to achieve better performance than the strategy of training the models on both corpora jointly. However, the observed improvement in model performance is not significant. For example, in Table 11, the XLM-R (Large) model improves the micro-averaged F1-score by only 0.8% when following the TL approach using DA compared to when the model is trained using only the augmented version of the Galén corpus. Meanwhile, in Table 11, the same model achieves an even smaller performance improvement, just 0.2%, when following the TL approach using DA compared to when trained using only the augmented version of the MEDDOCAN corpus. The results obtained demonstrate the difficulty of developing approaches that utilize the MEDDOCAN corpus to improve the performance of predictive models on the Galén corpus, and vice versa, once again highlighting the difference between the nature of the texts contained in these two corpora.

Finally, given the profound impact that generative language models (LLMs) on the field of NLP, we have initiated preliminary experiments to assess their efficacy in de-identification of unstructured clinical Spanish texts. It is important to acknowledge that recent studies have

demonstrated that GPT-style models, particularly in zero-shot configurations, are capable of attaining results that are comparable to those achieved by BERT-style models with minimal annotated data [71]. By employing GPT-4 [72], a SOTA reference model, in conjunction with three documents from the Galén corpus and a context prompt similar to that utilized in other de-identification studies [73,74], we have observed an overall satisfactory performance, both in terms of NER and document de-identification. The GPT-4 model demonstrated a 95% accuracy rate in recognizing all NEs containing personal information, 41 out of 43 NEs. However, the experiments exhibit significant limitations, particularly in accurately classifying NEs. Errors are prevalent, particularly in the categorization of locations that are classified as clinical centers or addresses, and in the categorization of personal identifiers, classified as medical record numbers. This behavior is similar to the observations made with Transformers. Additionally, the model incorrectly identifies clinical concepts, such as drug names and test numbers, as reference NE, despite the prompt indicating that these are merely reference numbers for analytical tests. The initial issue is inconsequential in the context of document de-identification, as it does not impede the efficient removal of personal information. However, the second issue has the potential to introduce a significant and undesirable consequence, removing crucial clinical concepts in the documents.

Furthermore, the experiment was conducted using a quantified version of the Starling-LM-7B model [75] obtained from Ollama. This lighter model, which lacks an Internet connection and is licensed under the Apache 2.0 open-source software license, could be integrated into a hospital computer system without compromising the system's integrity or privacy. Similar to the result obtained with GPT-4, Starling is able to recognize the majority of private information, in this case 32 out of 43. However, the aforementioned shortcomings are significantly exacerbated, resulting in a markedly inferior performance in NER and the redacting of a considerably larger amount of relevant information. To advance the study of generative LLMs in this task, it is essential to conduct a comprehensive analysis of the design of prompts, the utilization of diverse models, and the zero- and few-shot performance of the models, among other key aspects.

7. Conclusions

In this paper we have presented and analyzed the effectiveness of different NER models to address de-identification tasks on clinical domain corpora. On the one hand, we have analyzed the performance of different strategies based on RNNs with LSTM units, combined with CRF output layers. On the other hand, several Transformer-based models have been proposed for analysis in de-identification tasks. For this purpose, we have taken as a starting point Transformer-based models that have been pre-trained with both multilingual and Spanish corpora, general domain and clinical domain. Some of these Transformers have been adapted to the clinical domain by means of continual pre-training with a proprietary oncological clinical corpus, called Galén. The effectiveness of all these models of RNNs and Transformers has been tested on two corpora of clinical nature, one of them with labeled texts from clinical practice (Galén corpus) and the other consisting of a collection of clinical cases from the specific literature (MEDDOCAN corpus).

As a general conclusion, it is observed that Transformer models achieve higher performance rates than models based on RNNs, both for the F1-score and recall metrics. In particular, the XLM-R Large model, based on the RoBERTa model and pre-trained with a multilingual corpus, achieved the highest performance rates among all the models analyzed in this paper, thus demonstrating the performance of general-purpose and multilingual models in domain-specific text de-identification tasks, such as the clinical domain.

The different setups analyzed in this study also show that TL strategies based on training the models first on one corpus and then using those models to do inference on a different corpus (e.g. training models on Galen and using them to infer the entities present in MEDDOCAN) do not give good results, which leads to the conclusion that the specificities of the nature of each corpus clearly condition the effectiveness of the models in the de-identification tasks tacked here. On the other hand, the use of extended corpora by joining different corpora (Galén+MEDDOCAN) or by using data augmentation strategies through entity surrogation, has shown in this study a significant efficiency improvement in the inference capacity of the trained models, especially in the case of the Transformers. Furthermore, it has also been shown that TL strategies in which continual pre-training is carried out (e.g., pre-training a Transformer with the Galen corpus and subsequently fine-tuning this same model with the MEDDOCAN corpus) allows to obtain even better performance rates than training models with joint corpora. Nevertheless, the analysis shows that the enhanced performance is associated with a greater computational effort during training, attributable to the substantial number of trainable parameters. Although it has no impact on inference.

Finally, this paper has also shown the feasibility of incorporating the trained NER models for de-identification in the development of an end-user application for the management of the clinical personal information for the de-identification process in a real hospital environment. The incorporation of these models into clinical practice support tools contributes significantly to the efficient management of clinical information and, therefore, to the improvement of patient care in hospital environments.

To continue this work, validating the results on corpus from medical centers of other Spanish regions is essential. Furthermore, with the advent of generative AI, we aim to conduct a more comprehensive investigation into the potential applications of these models in the context of PHI detection and de-identification. Finally, it is intended to follow a boosting strategy through multiple cycles to improve the developed NLP models. In this way, un-identified available documents could be automatically labeled and de-identified using the NER models and de-identification tool presented at this paper. These paired labeled original and de-identified file require review by a human expert, which is less arduous work than labeling from scratch. After review, these documents can be incorporated into the corpus used to re-train the models, which intuitively should improve the performance of the NER models.

CRediT authorship contribution statement

Francisco J. Moreno-Barea: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Guillermo López-García:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Héctor Mesa:** Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nuria Ribelles:** Resources, Funding acquisition. **Emilio Alba:** Resources, Funding acquisition. **José M. Jerez:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Francisco J. Veredas:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

None Declared.

Acknowledgments

The authors acknowledge the support from the Ministerio de Ciencia e Innovación (MICINN) under project PID2020-116898RB-I00, from the Universidad de Málaga and Junta de Andalucía through grant UMA20-FEDERJA-045, from Pfizer S.L., the University of Malaga and the Fundación General UMA (UMA-FGUMA-Pfizer) through private funds.

References

[1] M. Douglass, G.D. Clifford, A. Reisner, G.B. Moody, R.G. Mark, Computer-assisted de-identification of free text in the MIMIC II database, in: Computers in Cardiology, 2004, IEEE, 2004, pp. 341–344, http://dx.doi.org/10.1109/CIC.2004.1442942.

[2] D.A. Dorr, W. Phillips, S. Phansalkar, S.A. Sims, J.F. Hurdle, Assessing the difficulty and time cost of de-identification in clinical narratives, Methods Inf. Med. 45 (03) (2006) 246–252, http://dx.doi.org/10.1055/s-0038-1634080.

[3] Act, Accountability, Health insurance portability and accountability act of 1996, Public Law 104 (1996) 191.

[4] Portability, Insurance and Act, Accountability, Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule, 2012.

[5] Council of the European Union, Regulation (EU) 2016/679 of the European parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General data protection regulation), Off. J. Eur. Union 119 (2016) 1–88.

[6] Cortes Generales de España, Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, Boletín Oficial Estado (2018).

[7] R. Chevrier, V. Foufi, C. Gaudet-Blavignac, A. Robert, C. Lovis, et al., Use and understanding of anonymization and de-identification in the biomedical literature: scoping review, J. Med. Internet Res. 21 (5) (2019) e13484, http://dx.doi.org/10.2196/13484.

[8] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (1) (2007) 3–26, http://dx.doi.org/10.1075/li.30.1.03nad.

[9] J. Guo, G. Xu, X. Cheng, H. Li, Named entity recognition in query, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 267–274, http://dx.doi.org/10.1145/1571941.1571989.

[10] B. Babych, A. Hartley, Improving machine translation quality with automatic named entity recognition, in: Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools, Resource and Tools for Building MT At EACL 2003, 2003.

[11] C. Aone, M.E. Okurowski, J. Gorlinsky, Trainable, scalable summarization using robust NLP and machine learning, in: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, 1998, pp. 62–66, http://dx.doi.org/10.3115/980845.980856.

[12] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Unsupervised named-entity extraction from the web: An experimental study, Artif. Intell. 165 (1) (2005) 91–134, http://dx.doi.org/10.1016/j.artint.2005.03.001.

[13] D. Mollá, M. Van Zaanen, D. Smith, Named entity recognition for question answering, in: Proceedings of the Australasian Language Technology Workshop 2006, 2006, pp. 51–58.

[14] R. Grishman, B.M. Sundheim, Message understanding conference-6: A brief history, in: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.

[15] L. Sweeney, Replacing personally-identifying information in medical records, the scrub system., in: Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 1996, p. 333.

[16] F.J. Friedlin, C.J. McDonald, A software tool for removing patient identifying information from clinical documents, J. Am. Med. Inform. Assoc. 15 (5) (2008) 601–610, http://dx.doi.org/10.1197/jamia.M2702.

[17] I. Neamatullah, M.M. Douglass, L.-W.H. Lehman, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, G.D. Clifford, Automated de-identification of free-text medical records, BMC Med. Inform. Decis. Mak. 8 (1) (2008) 1–17, http://dx.doi.org/10.1186/1472-6947-8-32.

[18] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106, http://dx.doi.org/10.1007/BF00116251.

[19] S.R. Eddy, Hidden markov models, Curr. Opin. Struct. Biol. 6 (3) (1996) 361–365, http://dx.doi.org/10.1016/S0959-440X(96)80056-X.

[20] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28, http://dx.doi.org/10.1109/5254.708428.

[21] H. Xue, S. Chen, Q. Yang, Structural support vector machine, in: International Symposium on Neural Networks, Springer, 2008, pp. 501–511, http://dx.doi.org/10.1007/978-3-540-87732-5_56.

[22] J.D. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 282–289, http://dx.doi.org/10.5555/645530.655813.

[23] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.

[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[25] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, 2015, arXiv preprint arXiv:1508.01991.

[26] J.P. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Trans. Assoc. Comput. Linguist. 4 (2016) 357–370, http://dx.doi.org/10.1162/tacl_a_00104.

[27] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016, arXiv preprint arXiv:1603.01354.

[28] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, 2016, arXiv preprint arXiv:1603.01360.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019, pp. 4171–4186.

[31] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240, http://dx.doi.org/10.1093/bioinformatics/btz682.

[32] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. (HEALTH) 3 (1) (2021) 1–23, http://dx.doi.org/10.1145/3458754.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451, http://dx.doi.org/10.18653/v1/2020.acl-main.747, Online.

[34] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Transformers for clinical coding in Spanish, IEEE Access 9 (2021) 72387–72397.

[35] C.P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199, http://dx.doi.org/10.18653/v1/2022.bionlp-1.19, URL https://aclanthology.org/2022.bionlp-1.19.

[36] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C.P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish Language Models, Procesamiento Lenguaje Natural 68 (2022) 39–60, http://dx.doi.org/10.26342/2022-68-3.

[37] F. Dernoncourt, J.Y. Lee, O. Uzuner, P. Szolovits, De-identification of patient notes with recurrent neural networks, J. Am. Med. Inform. Assoc. 24 (3) (2017) 596–606, http://dx.doi.org/10.1093/jamia/ocw156.

[38] G. López-García, F.J. Moreno-Barea, H. Mesa, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Named entity recognition for de-identifying real-world health records in Spanish, in: International Conference on Computational Science, Springer, 2023, pp. 228–242, http://dx.doi.org/10.1007/978-3-031-36024-4_17.

[39] N. Ribelles, J.M. Jerez, D. Urda, J.L. Subirats, A. Márquez, C. Quero, L. Franco, Galén: Sistema de Información para la gestión y coordinación de procesos en un servicio de Oncología, RevistaeSalud 6 (21) (2010) 1–12.

[40] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J.L. Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results, in: IberLEF@ SEPLN, 2019, pp. 618–638.

[41] H. Yang, J.M. Garibaldi, Automatic detection of protected health information from clinic narratives, J. Biomed. Inform. 58 (2015) S30–S38, http://dx.doi.org/10.1016/j.jbi.2015.06.015.

[42] J.Y. Lee, F. Dernoncourt, O. Uzuner, P. Szolovits, Feature-augmented neural networks for patient note de-identification, in: Proceedings of the Clinical Natural Language Processing Workshop, ClinicalNLP, 2016, pp. 17–22.

[43] Z. Liu, B. Tang, X. Wang, Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, J. Biomed. Inform. 75 (2017) S34–S42, http://dx.doi.org/10.1016/j.jbi.2017.05.023.

[44] Z. Jiang, C. Zhao, B. He, Y. Guan, J. Jiang, De-identification of medical records using conditional random fields and long short-term memory networks, J. Biomed. Inform. 75 (2017) S43–S53, http://dx.doi.org/10.1016/j.jbi.2017.10.003.

[45] K. Lee, M. Filannino, Ö. Uzuner, An empirical test of GRUs and deep contextualized word representations on de-identification, Stud. Health Technol. Inform. 264 (2019) 218–222, http://dx.doi.org/10.3233/shti190215.

[46] C. Grouin, A. Névéol, De-identification of clinical notes in French: towards a protocol for reference corpus development, J. Biomed. Inform. 50 (2014) 151–161, http://dx.doi.org/10.1016/j.jbi.2013.12.014.

[47] Z. Jian, X. Guo, S. Liu, H. Ma, S. Zhang, R. Zhang, J. Lei, A cascaded approach for Chinese clinical text de-identification with less annotation effort, J. Biomed. Inform. 73 (2017) 76–83, http://dx.doi.org/10.1016/j.jbi.2017.07.017.

[48] P. Richter-Pechanski, A. Amr, H.A. Katus, C. Dieterich, Deep learning approaches outperform conventional strategies in de-identification of German medical reports, in: GMDS, 2019, pp. 101–109, http://dx.doi.org/10.3233/SHTI190813.
[49] T. Jan, D. Trienschnigg, C. Seifert, D. Hiemstra, Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records, in: ACM Health Search and Data Mining Workshop, HSDM 2020, 2020.
[50] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in Spanish, Corpus, guidelines, methods and results, in: IberLEF@ SEPLN, 2020, pp. 303–323.
[51] R. Vunikili, H. Supriya, V.G. Marica, O. Farri, Clinical NER using Spanish BERT embeddings, in: IberLEF@ SEPLN, 2020, pp. 505–511.
[52] L. Akhtyamova, Named entity recognition in Spanish biomedical literature: Short review and bert model, in: 2020 26th Conference of Open Innovations Association, FRUCT, IEEE, 2020, pp. 1–7, http://dx.doi.org/10.23919/FRUCT48808.2020.9087359.
[53] L. Akhtyamova, P. Martínez, K. Verspoor, J. Cardiff, Testing contextualized word embeddings to improve NER in Spanish clinical case narratives, IEEE Access 8 (2020) 164717–164726, http://dx.doi.org/10.1109/ACCESS.2020.3018688.
[54] L. Lange, H. Adel, J. Strötgen, NLNDE: The neither-language-nor-domain-experts' way of Spanish medical document de-identification, 2020, arXiv preprint arXiv:2007.01030.
[55] N. Perez, L. García-Sardiña, M. Serras, A. Del Pozo, Vicomtech at MEDDOCAN: Medical document anonymization, in: IberLEF@ SEPLN, 2019, pp. 696–703.
[56] I. Pérez-Díez, R. Pérez-Moraga, A. López-Cerdán, J.-M. Salinas-Serrano, M.d. la Iglesia-Vayá, De-identifying Spanish medical texts-named entity recognition applied to radiology reports, J. Biomed. Semant. 12 (1) (2021) 1–13, http://dx.doi.org/10.1186/s13326-021-00236-2.
[57] R. Weegar, A. Pérez, A. Casillas, M. Oronoz, Recent advances in Swedish and Spanish medical entity recognition in clinical texts using deep neural approaches, BMC Med. Inform. Decis. Mak. 19 (2019) 1–14, http://dx.doi.org/10.1186/s12911-019-0981-y.
[58] S. Santiso, A. Casillas, A. Pérez, M. Oronoz, Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish, in: Bioinformatics and Biomedical Engineering: 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26–28, 2017, Proceedings, Part I 5, Springer, 2017, pp. 177–188, http://dx.doi.org/10.1007/978-3-319-56148-6_15.
[59] J. Koontz, M. Oronoz, A. Pérez, Evaluating data augmentation for medication identification in clinical notes, in: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 578–585.
[60] P. Báez, F. Villena, M. Rojas, M. Durán, J. Dunstan, The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, 2020, pp. 291–300, http://dx.doi.org/10.18653/v1/2020.clinicalnlp-1.32.
[61] M. Fernández, F. Villena, M. Rojas, F. Núñez, J.F. Silva, J. Dunstan, A transcription and information extraction system to facilitate EHR documentation in Spanish, 2023, http://dx.doi.org/10.21203/rs.3.rs-3175804/v1, Preprint.
[62] C. Aracena, L. Miranda, T. Vakili, F. Villena, T. Quiroga, F. Núñez-Torres, V. Rocco, J. Dunstan, A privacy-preserving corpus for occupational health in Spanish: Evaluation for NER and classification tasks, in: Proceedings of the 6th Clinical Natural Language Processing Workshop, 2024, pp. 111–121.
[63] D. Urda, N. Ribelles, J.L. Subirats, L. Franco, E. Alba, J.M. Jerez, Addressing critical issues in the development of an oncology information system, Int. J. Med. Inform. 82 (5) (2013) 398–407, http://dx.doi.org/10.1016/j.ijmedinf.2012.08.001.
[64] L.A. Ramshaw, M.P. Marcus, Text chunking using Transformation-Based learning, in: Natural Language Processing using Very Large Corpora, Springer Netherlands, Dordrecht, 1999, pp. 157–176.
[65] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Explainable clinical coding with in-domain adapted transformers, J. Biomed. Inform. 139 (2023) 104323, http://dx.doi.org/10.1016/j.jbi.2023.104323.
[66] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, http://dx.doi.org/10.48550/arXiv.1907.11692, arXiv [cs.CL].
[67] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Detection of Tumor Morphology Mentions in Clinical Reports in Spanish Using Transformers, in: Advances in Computational Intelligence, Springer International Publishing, Cham, 2021, pp. 24–35, http://dx.doi.org/10.1007/978-3-030-85030-2_3.
[68] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, L. Jorm, De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models, J. Biomed. Inform. 135 (2022) 104215, http://dx.doi.org/10.1016/j.jbi.2022.104215.
[69] P. López-Ubeda, M.C. Díaz-Galiano, L.A.U. López, M.T.M. Valdivia, Anonymization of clinical reports in Spanish: a hybrid method based on machine learning and rules, in: IberLEF@ SEPLN, 2019, pp. 687–695.
[70] A. Stubbs, C. Kotfila, Ö. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, J. Biomed. Inform. 58 (2015) S11–S19, http://dx.doi.org/10.1016/j.jbi.2015.06.007.
[71] Á. García-Barragán, A. González Calatayud, O. Solarte-Pabón, M. Provencio, E. Menasalvas, V. Robles, GPT for medical entity recognition in Spanish, Multimedia Tools Appl. (2024) 1–20, http://dx.doi.org/10.1007/s11042-024-19209-5.
[72] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, 2023, arXiv preprint arXiv:2303.08774.
[73] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, et al., Deid-gpt: Zero-shot medical text de-identification by gpt-4, 2023, arXiv preprint arXiv:2303.11032.
[74] J.A. Lund, K.Ø. Mikalsen, J. Burman, A.Z. Woldaregay, R. Jenssen, Instruction-guided deidentification with synthetic test cases for norwegian clinical text, in: Northern Lights Deep Learning Conference, PMLR, 2024, pp. 145–152.
[75] B. Zhu, E. Frick, T. Wu, H. Zhu, J. Jiao, Starling-7b: Improving llm helpfulness & harmlessness with rlaif, 2023.