Applied Soft Computing 145 (2023) 110540

Contents lists available at ScienceDirect

Applied Soft Computing

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

AGORA: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents

Rodrigo Juez-Hernandez [a], Lara Quijano-Sánchez [a,b,*], Federico Liberatore [c,b], Jesús Gómez [d]

[a] Escuela Politécnica Superior, Universidad Autonóma de Madrid, Madrid, Spain
[b] UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Madrid, Spain
[c] School of Computer Science & Informatics, Cardiff University, UK
[d] Oficina Nacional de Lucha Contra los Delitos de Odio, Ministerio del Interior, Madrid, Spain

ARTICLE INFO

Article history:
Received 16 January 2023
Received in revised form 24 May 2023
Accepted 13 June 2023
Available online 27 June 2023

Keywords:
Document anonymization
Information extraction
Named Entity Recognition
Natural language processing
Visualization tools
Document sharing

ABSTRACT

Public institutions, such as law enforcement agencies or health centers, have a vast volume of unstructured text documents, e.g. police reports. Currently, before this data can be shared (e.g. with research institutions), it must go through a lengthy and costly human anonymization procedure.

This paper addresses this issue by presenting AGORA, a cutting-edge tool that automatically identifies key entities and anonymizes sensitive data in text documents. AGORA has been developed in partnership with the Spanish National Office Against Hate Crimes and validated in the police and medical domains. This tool allows to export both anonymized texts and identified entities to structured files, thus, simplifying its exploitation for analysis purposes. Also, AGORA is capable of plotting the location entities identified in the documents, as well as obtaining and displaying relevant information from their geographical surroundings. Thus, it simplifies the task of generating comprehensive datasets for subsequent data analysis or predictive tasks. Its main goal is to foster cooperation between public institutions and research centers by easing document sharing as well as serving as a foundation for addressing succeeding phases in data science.

The paper conducts a comprehensive assessment of the literature on Named Entity Recognition methodologies and technologies. Followed by extensive computational experiments to identify the best configuration for the NER models embedded in AGORA which include both successful state-of-the-art model setups and novelly proposed ones. Finally, the methodology, conclusions and software provided can be easily reused in similar application scenarios.

1. Introduction

In the last decades, society has experienced radical structural changes, mainly due to technological evolution [1]. In particular, in the field of citizen security or medicine, Artificial Intelligence (AI) may represent a wide range of opportunities to improve prevention and response to various events, as well as challenges to overcome. Currently, significant efforts are being made in both digitization and the use of available data [2–4], where, there are numerous examples of public institutions adapting their structures and operating processes in response to changes and new technologies such as AI [5–8]. While its use is controversial in many ways, the main obstacles are related to people's privacy [9,10], or to the potential biases that the new algorithms can generate, which can affect citizens equality [11,12]. Given that personal data protection refers to a collection of legal and computer procedures aimed at ensuring individuals' rights to control their personal data [13], there may be a potential conflict between data security and research in these areas.

Collaborations between public institutions and research centers have previously been possible thanks to manual document anonymization. This data treatment approach, combined with Non-Disclosure Agreements, ensures a high level of security. However, this procedure is time consuming and expensive. For instance, one annotator reading 20.000 words per hour is projected to cost around 50 cents per hour [14]. Consequently, a tool that speeds up this procedure might save a lot of money and time while also improving research collaborations. For example, when researchers at a medical facility were asked if they would share clinical data for study [15], only about 53% of respondents

* Correspondence to: Escuela Politécnica Superior, Universidad Autónoma de Madrid, C/ Francisco Tomás y Valiente, 11, Campus de Cantoblanco, 28049 Madrid, Spain.

E-mail addresses: rodrigo.juezh@estudiante.uam.es (R. Juez-Hernandez), lara.quijano@uam.es (L. Quijano-Sánchez), liberatoreF@cardiff.ac.uk (F. Liberatore), ses.ondod@interior.es (J. Gómez).

https://doi.org/10.1016/j.asoc.2023.110540

said yes. However, acceptance increased to a startling 90% when the same topic was given again, but this time anonymization was guaranteed on such data.

Nevertheless, anonymization is not the only desirable goal. A tool that incorporates information retrieval tasks might contribute to enrich research projects noting that the aforementioned AI approaches are primarily data driven and, that these free text documents hold much of unexplored information that could be automatically processed. As a result, it would be advantageous to infer relevant data and upload it to institutional databases, as it could be used as explanatory factors in various predictive systems. For example, many investigations require knowledge of the spatial locations (i.e. geolocations) of events of interest, since they provide useful information [16].

Finally, the people who enable and exchange said information in multidisciplinary projects, where public institutions' efforts and data are merged with cutting-edge AI technology, do not necessarily know how to extract and manipulate files and data. Consequently, it would be ideal for such tool to provide an easy-to-use interface where people without specific data science skills may analyze and examine the relevant information in documents.

In this paper, we present a tool for document anonymization, geoparsing, geocoding, and visualization which addresses the above real world necessities. For its development, the following research questions were formulated and answered:

RQ1: What is the state of the art in anonymization, geoparsing and geocoding of written documents? Our review showed that all these tasks require a Named Entity Recognition (NER) model. Also, we realized that most works are in the medical domain and in the English language.

RQ2: Is it necessary to train ad hoc models or can we use pretrained models, independently of the domain? Our experiments indicate that ad hoc models perform much better.

RQ3: What is the best architecture for ad hoc models? Our analysis points out to the FlairNLP framework, using a Bidirectional LSTM or GRU network with a CRF classifier combined with advanced embeddings.

RQ4: Are trained ad hoc models and their architectures portable across domains? Our cross-domain investigation shows that although they could be employed there is a significant performance gap.

In summary, the contributions of this paper are many-fold:

1. The first intelligent tool that combines document anony mization, geoparsing, geocoding and visualization and is domain-agnostic.
2. An extensive analytical literature review on NER, anony mization, geoparsing, and geocoding.
3. The first anonymization model in the police context.
4. A methodical comparison of Tensorflow, FlairNLP, Transformers, and pretrained models in NER which addresses a gap in the literature.
5. A cross-domain evaluation of NER models in the Spanish language.

Additionally, the present research resulted in the following outputs: an open-source intelligent tool, AGORA (Analysis and Geolocalization Of Reports Anonymized),[1] public libraries for benchmarking[2] and tagging,[3] and a NER model in the Spanish language for the medical context.[4] AGORA, has been developed in collaboration with and is currently in use by the Spanish National Office Against Hate Crimes (ONDOD).

The remainder of this paper is structured as follows: Section 2 provides a summary of current research in NER, anonymization, geocoding, geoparsing, and other similar developed tools. Section 3 describes our NER research and the models to be evaluated. Next, the datasets used in the studied domains (police and medical) are presented in Section 4. Section 5 focuses on the computational experiments and their outcomes. Section 6 contains a description of the AGORA tool. Section 7 highlights the research findings and concludes the paper.

2. Background and related works

This section presents an overview of the state of the art in NER based models and tools for anonymization, geoparsing, and geocoding. For this, we have conducted a thorough bibliographic review where the above keywords have been combined with additional domain-related terms to construct several queries used to search various academic and bibliographic databases (i.e. ACM, IEES, WOS and SCOPUS).

2.1. The problem of Named Entity Recognition

NER is a subtask of information extraction that aims to discover and classify named entities mentioned in unstructured text into predefined categories such as person names, organizations, locations, time expressions, etc [17]. It is used for a variety of tasks; in our case, it is employed for geographical data extraction and anonymization. The work of Karimzadeh et al. [18] and Schmitt et al. [19] in this field (summarized in Table 1) stands out as it examines the state of the art in NER training models, commercial tools, and available datasets. They conclude that: (i) most NER tools include a base version that has been pretrained using a dataset, and that users can tailor it to their own workflow by using their own corpus; (ii) many of them have more than one base model and are available in multiple languages; and (iii) one of the most pressing challenges in the field is the scarcity of datasets in languages other than English. As a result, systems that do propose linguistic options fall behind their English counterparts.

2.1.1. Languages other than English

Given that AGORA has been developed in collaboration with and for the use of Spanish institutions, we further study NER approaches for languages other than English. Table 2 summarizes the research papers found.

2.1.2. NER embeddings

Embeddings, which represent words as numeric vectors, play an important role in NER research. Many of the developed models (such as neural networks) rely on them because they cannot accept words as input and instead require vectors. GloVe, Word2Vec, MUSE, FastText, Flair, Cross-lingual embeddings, BPEMb, mBERT, MADE, CBOW, CONEC GLOBAL & CONEC GLOBAL are examples of NER state-of-the-art identified embeddings.

[1] Application to be released upon paper acceptance: https://github.com/rjuez00/AGORA.

[2] Available at: https://github.com/rjuez00/nereval, last accessed May 24, 2023.

[3] Available at: https://github.com/rjuez00/doccano-transformer.

[4] Available at: https://huggingface.co/rjuez00, last accessed May 24, 2023.

Table 1
Most common NER tools as of [18] and [19].

NER Tool	Architecture	Corpus for base version	Other languages
Stanford NLP [20]	CRF	CoNLL2003, Ritter, MSM2013	German, Spanish, Chinese
NLTK [21]	Max Entropy Model	Stanford NER	German, Spanish, Chinese
SpaCy [22]	CNNs (2.0), Transformers (3.0)	Wikigold, OntoNotes	Catalan, Chinese, Danish, Dutch, Finnish, French, German, Greek, Italian, Japanese, Korean, Lithuanian, Macedonian, Norwegian, Polish, Portugese, Romanian, Russian, Spanish, Swedish
CogComp [23]	Perceptron	CoNLL2003, Enron email	NO
OpenNLP [24]	Maximum Entropy Model	CoNLL2003	Danish, German, Spanish, Portuguese, Swedish, Dutch
LingPipe [25]	Hidden Markov Models	MUC-6	Arabic, Chinese, Dutch, German, Greek, Hindi, Japanese, Korean, Portuguese and Spanish
GATE [26]	Rule Based	N/A	NO
MIT IE [27]	SVM	CoNLL2003, ACE2002, Wikipedia, Freebase, GigaWord	Spanish, and German
TextRazor [28]	N/A	MSM2013, Ritter, UMBC	Arabic, Chinese, Danish, Dutch, Finnish, French, German, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish
Flair [29]	BiLSTM, Transformers	CoNLL2003, OntoNotes	Arabic, German, French, Spanish, Dutch, Danish, Malayalam, Portuguese.

Table 2
State of the art summary for NER in other languages papers.

Paper	Language	Domain	Best model	Results
[30]	Spanish	Legal	SpaCy3 fine tune/ReGex and Gazeteers	F1 0.9316, Recall 0.923, Precision 0.9666
[31]	Russian	News	CRF++	F1 0.8993
[32]	Arabic	News and Webpages	Neural Network 1 Hidden Layer	F1 0.8864
[33]	Indonesian	Tweets	Hidden Markov Models, word embeddings	F1 0.7561, Recall 0.5568
[34]	Albanian	Webpages.	Word-Char Embeddings BiLSTM+CRF	Precision 0.763, Recall 0.7459, F1 0.7543
[35]	Indonesian	CNN Indonesian News.	Word2Vec + BiLSTM	Recall 0.8532, Precision 0.8114, F1 0.8318
[36]	Chinese	Typhoon reporting Tweets	BiLSTM	Precision 0.42, Recall 0.61, F1 0.5
[37]	Italian	General	POS rules to detect entities using MapReduce	Recall 0.9, Precision 0.87 and F1 0.76

2.1.3. NER-based tools

Given that AGORA aims to enhance multidisciplinary collaborations and provide and easy-to-use interface we have also studied the state of the art focused on bringing ad hoc NER models closer to services and the related UI tools that people without technical knowledge can use. Table 1 summarizes the research papers found.

2.2. Anonymization

This section describes the identified works related to text anonymization. For this task, NER is used in clinical reports, messaging, and legal datasets to detect and remove sensitive information from documents. The most common techniques usually involve a Recurrent Neural Network like LSTM or GRU trained ad hoc along domain-specific rules. To illustrate this, Chomutare [38] conducts a state of the art review of projects that use the i2b2 public clinical dataset [39] for anonymization, where only two of the nine publications agree on the amount of sensitive information to anonymize in the dataset. The most commonly used model in the studies is BiLSTM+CRF. This review includes the work by Ahmed et al. [40], that obtain their best precision of 0.9901 when using a Bidirectional GRU. The additional reviewed research papers are summarized in Table 3.

2.3. Geoparsing and geocoding

This section explores existing geoparsing and geocoding research. Geoparsing is the process of extracting locations from text utilizing techniques such as NER, ReGex rules, and Gazetteers. Since the purpose of recognizing entities in a text is similar to that of anonymization, the models, embeddings, and procedures observed are very similar. It is important to highlight that geocoding and geoparsing are not the same thing although related, i.e. after extracting locations from text (geoparsing), researchers unravel their true location on a map (geocoding), meaning that they associate the extracted words with a latitude and a longitude. Table 4 provides an overview of the reviewed research papers.

2.4. Limitations and research opportunities

Tables 1, 2, 3 and 4 provide an overview of the publications and methodologies revised. In summary we see that English based NER models typically outperform their counterparts in other languages, with the lowest and highest F1 being 0.82855 and 0.9828. Meanwhile, in Spanish, the same techniques yield an F1 of 0.90381–0.9810, though the average is lower.

After reviewing the literature (**RQ1**), we identify the following shortcomings and gaps: All the models proposed in the literature are ad hoc and are not tested against others. Most works address the problem in the medical context; also, the presence of works focusing on languages other than English is very limited. Furthermore, we detected that pretrained models are completely disregarded (**RQ2**). More in detail on ad hoc models, heterogeneity in terms of models is very limited. Most contributions combine a Bi-LSTM model with Word2Vec, Flair, or character embeddings, appearing in 16 of the projects reviewed. The BiGRU model, on

Table 3
State of the art summary for anonymization papers.

Paper	Language	Domain	Best model	Results
[40]	English	Medical	BiGRU	Precision 0.9901, Recall 0.9841, F1 0.9822
[41]	English	Medical	BiLSTM+CRF	F1 0.9828 no type, 0.9698 token match
[42]	Spanish	Medical	Glove+Char+BiLSTM+CRF	Precision 0.99 Recall 0.97, F1 0.98
[14]	English	General	Encoder–decoder	Recall 0.9891, Precision 0.9812
[43]	English	Medical	ELECTRA	F1 0.9863
[44]	English	Medical	GloVe + Flair + BiLSTM+CRF	F1 0.9614
[45]	Portuguese	Medical	Word2Vec + Flair(Portuguese)+BiLSTM+CRF	Precision 0.9256, Recall 0.9578, F1 0.9413
[46]	English	Medical	BiLstm+CRF	Precision 0.8391, Recall 0.818
[47]	English	Facebook Messages	BiLSTM	Precision 0.2992, Recall 0.865
[48]	Italian	Medical	Flair+FastText+BiLSTM+CRF	Precision 0.7953
[49]	Spanish	Medical	ReGex and CRF	Recall 0.9567
[50]	Spanish	Medical	ReGex + CRF + SVM	Precision 0.92113, Recall 0.888712, F1 0.90381
[51]	French	Medical	Unitex (rule based)	Precision 1.00, Recall 0.98
[52]	German	Medical	ReGex, Gazetteers, Levenshtein distance	Precision 0.8, Recall 0.78, F1 0.78.
[53]	German	Medical	BiLSTM+CRF	Precision 0.97, Recall 0.955
[54]	Italian	Medical	MultiBPEmb + Flair + BiLSTM+CRF trained with English and Italian dataset	F1 0.9449
[55]	English	Medical	CommonCrawl Embeddings + BiLSTM+CRF	F1 0.9584
[56]	English	Medical	Skip-gram word embedding model	Recall 0.8124
[57]	English	Medical	BiLSTM+CRF	Recall 0.9925, Precision 0.9921
[58]	German	Email	BiLSTM+CRF	Precision 0.9368 Recall 0.8727
[59]	English	Medical	Dictionaries, heuristics, NLP	Precision 0.99-1.0-0.8-0.92 (patient-doctors-other,places)
[60]	English	Medical	LSTM Real data + synthetic	Recall 0.879, F1 0.901
[61]	English	Medical	CharEmbeddings +BiLSTM+CRF	Precision 0.9202, Recall 0.8404
[62]	French	Medical	ReGex + CRF + Xerox Incremental Parser	Recall 0.957, F1 0.97

Table 4
State of the art summary for geoparsing and geocoding papers.

Paper	Language	Domain	Best model	Results
[63]	English	News	Hand Crafted Rules + Domain Specific ML	F1 0.7213, Accuracy 0.73
[64]	French	Housing Advertisements.	Flair+CamemBERT+Word2Vec + BiLSTM+CRF	Precision 0.863, Recall 0.889, F10.876
[65]	Spanish	News	ConEc Embeddings + 1 hidden layer Neural Network	Accuracy 0.9633, Precision 0.7085, Recall 0.5761, F1 0.893
[66]	French	19th century Novels.	"PERDIDO" platform	Precision 0.997, Recall 0.99, F1 0.893 0.993
[67]	English/Turkish	Social Media	map-database entity matching (OpenStreetMaps gazetteer match)	Precision 0.96–0.99, F1 0.90–0.97
[68]	Spanish	Geography	Perceptron with 3 hidden units	F1 0.8093, Recall 0.5663
[69]	German	Tweets	Linear-chain CRF	Recall 0.504, Standford NER precision 0.79

the other hand, is completely overlooked. In terms of frameworks, there is a great predominance of Tensorflow. The FlairNLP and the Transformer frameworks are used only in one paper each. Therefore, by studying the literature alone it is not possible to discern what configuration provides the best results for ad hoc models (**RQ3**). Finally, to the best of the authors knowledge, there are no contributions that look at the robustness of the models in terms of inter-domain transferability (**RQ4**).

3. Methodology

In this section we illustrate the model configurations (embeddings, model, and architecture) and hyperparameters considered in this research. In order to select the best feasible NER model for our designed tool, we consider the best performing state-of-the-art models and either adopt their normal configuration, optimize them via hyperparameter tuning, or propose modifications to their setup. More details are provided in each of the following subsections, which introduce the implemented models.

3.1. Pretrained models

We studied Stanza[5] and Flair [29]. Other pretrained models, such as SpaCy, are not included in this research because it is well known that they perform significantly worse on the Spanish language, as demonstrated in [70] To test the performance of pretrained models we map the standard NER entities (i.e., PER and LOC, standing for "person" and "location") to our own domains' entities.

5 https://stanfordnlp.github.io/stanza/ner.html

3.2. Tensorflow framework

In Pérez-Díez et al. [42] authors argue that the three LSTM models they study yield the best results in Spanish. We studied the same models in our domains of interest adding hyperparameter tuning. Hence, we implemented the following Tensorflow state-of-the-art NER models: [TENS1] Huang et al. [71] Word GloVe Embeddings + LSTM + CRF classifier; [TENS2] Lample et al. [72] Character Embeddings stacked with Word GloVe Embeddings + LSTM + LSTM + CRF classifier; [TENS3] Ma and Hovy [73] Character Embeddings stacked with Word GloVe Embeddings + Convolutional Network + LSTM + CRF classifier; and compared them as in Pérez-Díez et al. [42]. However, while Pérez-Díez et al. [42] simply use the models' default parameters we compare the following configurations: 25 against 40 epochs, 0.1 against 0.2 learning rate, and batch size of four against eight. Preliminary experiments showed that the best hyperparameters configuration is: 25 epochs, learning rate of 0.1, and a batch size of 4.[6] Also, the hyperparameters specific to each model are detailed in the following:

- TENS1: LSTM size = $\{100, 200, 300\}$. Embeddings size = 300 (fixed, cannot be changed).
- TENS2 & TENS3: LSTM size = $\{100, 200\}$. Embeddings size = $\{100, 200\}$.

3.3. FlairNLP framework

FlairNLP [29] is a framework that includes a full set of pre-trained embeddings, including the Flair Embeddings. The library also allows the user to customize a model by allowing to select between LSTM, GRU and RNN cells, and whether or not to output to a CRF classifier. This framework has a superior memory management when compared to Tensorflow manual implementations. So far in the literature, few authors Patel [74] have tried this framework for NER models. In this research we study the following architectures, where the last three are novel to NER research: [FLAI1] Patel [74] BiLSTM+CRF with Spanish FastText Word Embeddings + Flair Embeddings; [FLAI2] BiLSTM+CRF with Spanish FastText Word Embeddings + BytePair Embeddings; [FLAI3] BiLSTM+CRF with Spanish FastText Word Embeddings + "bert-base-multilingual-cased" Fine-tuned Transformer Embeddings; [FLAI4] BiGRU+CRF with Spanish FastText Word Embeddings + Flair Embeddings.

The embeddings chosen are available in FlairNLP and trained for the Spanish language. The last configuration in the list uses GRU cells instead of LSTM; it is important to notice that this comparison is novel to NER models and not common in the state of the art, only proposed in Ahmed et al. [40]. Flair embeddings are used with that cell because they are the best performing in the LSTM tests.

We tested the following hyperparameters values: maximum epochs = 40; patience equal = 10; starting learning rates = $\{0.1, 0.2\}$; LSTM size = $\{128, 256\}$; RNN depth layers = $\{1, 2\}$. The batch size is fixed depending on the type of embeddings used: 3 for mBERT embeddings, 10 for Byte embeddings, and 6 for Flair embeddings.[7]

[6] Other configurations where tested, however, exceeded memory and computational time limits (32 GB and 1 week, for hyperparameter configuration).

[7] Other configurations where tested, however, exceeded memory and computational time limits (32 GB and 1 week, for hyperparameter configuration).

3.4. HuggingFace's transformer framework

Given the outstanding success of using transformers in other NLP tasks we also analyze their performance in the NER task, which had previously only been attempted by Ahmed et al. [46]. Specifically we use the Huggingface's Transformers library[8] to fine tune and test different Spanish and Multilingual models. This framework is designed to use each model "as is" and already includes an embedding layer (typical of Transformer models) so our research is focused on comparing the main available architectures instead of optimizing them. We novelly [tested] in the NER domain four different architectures, three for fine tuning and one for transfer learning: [TRAN1] RoBERTa transformer [75]; [TRAN2] BETO transformer [76]; [TRAN3] BERT multilingual [77]; [TRAN4] Fine tuned NER-C [76] transformer model.[9] Differently from the previous three, this model is pretrained. Therefore, to perform transfer learning, the output classifier is replaced by a linear Pytorch layer.

4. Datasets

The experimental evaluation of the above-mentioned models in the police and medical domains is carried out using the two datasets described below.

4.1. Police report dataset

For the police domain we have manually annotated a police report corpus. The dataset, which was provided by ONDOD, contains reports sourced randomly from Spanish police stations across the whole territory, dating back to 2013. The final corpus is comprised of 319 reports of the Spanish National Police, having a size comparable to that of other corpuses in the literature [48].

For the scope of this research, the annotated entity classes of interest are: "Person_Name", "Street_Name" (i.e., squares, streets, avenues, and other directions that can be unequivocally geocoded) and "General_Location" (i.e., cities, regions and other geographic areas). Differentiating between street names and general locations allows for fine-grained geolocation. The dataset also includes tags for "Temporal References", "Telephone Numbers", "Police Dependencies", "Identification IDs", "Police Diligence and Report IDs", and "Emails and Webpages". However, given that they are easy to identify with ReGex rules, these are not considered for the evaluation of the NER models' performance.

Each document is tagged by one person and validated by another. Since annotations may include leading or trailing empty spaces, commas, dots, or other characters that are not actually part of the entity, these are automatically cleaned using ReGex rules. To simplify development each entity can only have a unique tag.

4.2. MEDDOCAN dataset

In this section, we outline the dataset used for the medical domain, the MEDDOCAN corpus, whose details can be found in [78]. This dataset, used for the MEDDOCAN anonymization contest [78], was professionally created by PlanTL-GOB-ES. It is comprised of 1875 clinical cases and includes a set of 29 entity types.

[8] https://huggingface.co/docs/transformers/index

[9] This self-fine tuned BETO model for NER can be found at https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner.

Table 5
Test performance results for pretrained model Flair.

(a) Complete coverage

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.5672	0.7365	0.6386	0.7480
LOC	0.4523	0.5420	0.4270	
PER	0.7821	0.9309	0.8501	

(b) Partial coverage.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.5892	0.7663	0.6638	0.7783
LOC	0.3695	0.5686	0.4479	
PER	0.8088	0.9640	0.8796	

Table 6
Test performance results for pretrained model Stanza.

(a) Complete coverage.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.4908	0.65	0.5593	0.6747
LOC	0.3316	0.4464	0.3805	
PER	0.6499	0.8536	0.7380	

(b) Partial coverage.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.5103	0.6764	0.5917	0.7033
LOC	0.3492	0.4704	0.4008	
PER	0.6713	0.8824	0.7625	

5. Experimental results

This section presents the results of our methodology and derives the answers to the posed research questions.

5.1. Police domain

The results of the experiments on the *Police Report Dataset* allows us to answer research questions **RQ2** and **RQ3**. For this analysis, the dataset is split in 250/50/19 for training/testing/valida tion purposes. The total performance of each system is assessed using the F1 score. However, for each entity type we employ the most relevant metric for evaluating the performance, as explained in the following:

- Recall is used for the entity type "PERSON_NAME", since it shows how many person names are covered, which is essential for anonymization.
- Precision is the measure of choice for the entity type "STREET_NAME", since the geolocalization functionality needs real street names, and precision shows how many of the projected streets are correct.
- F1 is used for "GENERAL_LOCATION" as it provides a balance between recall and precision.

The measure "Anonymization Coverage" is also taken into account which is equivalent to the recall obtained disregarding the entities' type.[10]

The performance scores have been calculated using the tailor-made "nereval" library. This library takes into account both complete and partial coverage.

5.1.1. Pretrained models

Pretrained models make use of generic labels, i.e., LOC (location), PER (person), MISC (miscellaneous), ORG (organization). Therefore, for the sake of comparison and evaluation, it is necessary to perform an entity type mapping:

- (STREET_NAME, GENERAL_LOCATION) $\mapsto$ LOC
- PERSON_NAME $\mapsto$ PER

Tables 5(a), 5(b), 6(a) and 6(b) illustrate the performance for the pretrained models computed using complete and partial coverage, respectively. In the tables, columns "Complete" refer to complete coverage (the predicted entity must exactly match the real entity), while columns "Partial" refer to partial coverage (i.e., the predicted entity must match at least 80% of the real entity). Overall, the Spanish NER model Flair outperforms the Stanza (StandfordNLP) model.

By comparing these results with those of ad hoc models (see following subsections) it is possible to notice that the performance of pretrained models is significantly lower. However, the results for "PER" are quite impressive; in fact, the recall of 0.964 achieved by Flair's partial coverage rivals custom ad hoc models using Flair Embeddings. However, the latter provides better performances on all the entities, and achieves the same results for complete entity matching.

A clear issue with these tools is that recall is usually excellent, but precision (especially for "LOC") rarely exceeds 0.37. Both models clearly overpredict (i.e., identify too many entities). This can be a major issue because our tool's goal is to be able to use these texts for other tasks, so models that overpredict, delete too much information. Moreover the anonymization coverage is far from perfect, with more than 23% of entities leaking in the best-case scenario.

These findings are consistent with those of Chen et al. [36], who compare the Standford NER model to an ad hoc trained model, with the latter achieving significantly better results.

Table 7
Test scores of best model for TENS1. Best hyperparameter configuration: LSTM size = 200.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.7722	0.6643	**0.7142**	0.8695
STREET_NAME	**0.6466**	0.7107	0.6772	
GENERAL_LOCATION	0.7723	0.5096	0.6140	
PERSON_NAME	0.8071	**0.7564**	0.7809	

Table 8
Test scores of best model for TENS2. Best hyperparameter configuration: (LSTM $\times$ CHAR size) = (100 $\times$ 100).

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9163	0.9092	**0.9128**	0.9743
STREET_NAME	**0.8560**	0.8595	0.8577	
GENERAL_LOCATION	0.8954	0.8822	**0.8887**	
PERSON_NAME	0.9447	**0.9391**	0.9419	

5.1.2. Results Tensorflow

In this section, we present the results of running the models described in Section 3.2 on the *Police Report Dataset* and compare them to the results obtained in [42].

[10] For example, if we have the text "My name is Rodrigo Juez, I live in Street Bergantín", and the entities to remove are the name and the street, if the model detects both entities but misclassifies the name as a location, it will still have a 100% anonymization coverage.

Table 9
Test scores of best model for TENS3. Best hyperparameter configuration: (LSTM × CHAR size) = (100 × 200).

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9251	0.9051	**0.915**	0.9676
STREET_NAME	0.8247	0.8554	0.8398	
GENERAL_LOCATION	0.9120	0.8704	**0.8907**	
PERSON_NAME	0.9589	0.9401	0.9494	

Table 10
Test scores of best model for FLAI1. Best hyperparameter configuration: (WIDTH × DEPTH × LR) = (256 × 1 × 0.2).

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9266	0.9274	**0.927**	0.9844
STREET_NAME	**0.9174**	0.9136	0.9155	
GENERAL_LOCATION	0.8955	0.8968	0.8962	
PERSON_NAME	0.967	0.9717	0.9693	

Table 11
Test scores of best model for FLAI2. Best hyperparameter configuration: (WIDTH × DEPTH × LR) = (256 × 1 × 0.2).

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9086	0.906	**0.9072**	0.9771
STREET_NAME	0.8816	0.8926	0.8871	
GENERAL_LOCATION	0.8925	0.8821	**0.8873**	
PERSON_NAME	0.9516	0.9432	0.9474	

Table 12
Test scores of best model for FLAI3. Best hyperparameter configuration: (WIDTH × DEPTH × LR) = (256 × 1 × 0.2).

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9173	0.9266	**0.9219**	0.9808
STREET_NAME	0.9184	0.9259	0.9221	
GENERAL_LOCATION	0.8779	0.8881	**0.8829**	
PERSON_NAME	0.9556	0.9659	0.9607	

Table 13
Test scores of best model for FLAI4. Best hyperparameter configuration: (WIDTH × DEPTH × LR) = (128 × 2 × 0.1).

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9251	0.9338	**0.9294**	0.9875
STREET_NAME	0.8988	0.9136	0.9061	
GENERAL_LOCATION	0.9074	0.9113	**0.9094**	
PERSON_NAME	0.969	**0.9766**	0.9728	

Tables 7–9 present the scores of the best performing configurations for each Tensorflow model. Differently from the results reported in [42], which conclude that the best model is TENS2, our best F1 scoring model is TENS3. In our experiments, TENS2 and TENS3 perform very similarly, as is the case for [42]; It is worth to notice that both TENS2 and TENS3 are more complex than TENS1, which might explain why the former outperform the latter. Pérez-Díez et al. [42] report similar findings as the performance of TENS2 and TENS3 is fairly close. The fact that the models' performance in their domain and ours is quite similar suggests that Recurrent Neural Networks are quite portable.

5.1.3. Results FlairNLP

In this section we present the results of running the models described in Section 3.3 on the *Police Report Dataset*. A summary of the configuration that yields the best results for each model tested is shown in Tables 10–13.

Table 14
Scores of best results for TRAN1.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.876708	0.907372	0.891745	0.982140
STREET_NAME	0.823529	0.864198	0.843373	
GENERAL_LOCATION	0.873773	0.906841	0.890000	
PERSON_NAME	0.932821	0.951076	0.941860	

Table 15
Scores of best results for TRAN2.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.887388	0.918630	**0.902690**	0.984575
STREET_NAME	0.832031	0.876543	0.853707	
GENERAL_LOCATION	0.884993	0.918486	**0.901429**	
PERSON_NAME	0.945140	**0.960861**	0.952935	

Table 16
Scores of best results for TRAN3.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.878428	0.908693	0.893141	0.976944
STREET_NAME	0.831418	0.893004	0.861111	
GENERAL_LOCATION	0.870922	0.893741	0.882184	
PERSON_NAME	0.932945	0.939335	0.936129	

Table 17
Scores of best results for TRAN4.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.888996	0.916322	0.902431	0.979867
STREET_NAME	0.854331	0.893004	0.873239	
GENERAL_LOCATION	0.886202	0.906841	0.896403	
PERSON_NAME	0.926457	0.949119	0.937651	

The only other work performed in this framework is that by Patel [74] that achieve an F1 of 0.9017 on model FLAI1, without any hyperparameter tuning. For the same model, we report higher scores when applying it to our domain and performing hyperparameter tuning. This suggests that Flair embeddings are domain portable.

According to our experiments, the best F1 is achieved by model FLAI4 which replaces LSTM cells by GRU cells. The best anonymization coverage is also achieved by model FLAI4. It is worth noting that all LSTM models achieve their best performance with a width of 256, while GRU performs better with a smaller width of 128. Flair embeddings clearly outperform other configurations as seen when comparing the LSTM models (i.e., FLAI1, FLAI2, and FLAI3).

5.1.4. Results HuggingFace's transformers

In this section, we present the results of running the models described in Section 3.4 on the *Police Report Dataset* (see Tables 14–17). TRAN2 is the best performing model in every task, except for "STREET_NAME". The best result in "STREET_NAME" is TRAN4. Both models use BETO, although TRAN2 is trained specifically on the dataset while TRAN4 uses transfer learning.

5.1.5. RegEx extracted entities

As mentioned in Section 4.1. Some entities are extracted using RegEx filters. The precision of the RegEx has been tested with the same experimental configuration described above and the results are briefly given in the following: "Temporal References", 0.9146; "Telephone Numbers", 0.963; "Police Dependencies", 0.9688; "Identification IDs", 0.9375; "Police Diligence and Report IDs", 0.926; "Emails and Webpages", 1.0.

5.1.6. Discussion

In the previous section we have provided the results of testing several cutting-edge models with fine-tuning variations for our police domain in order to select the best for our tool. This allows us to provide an answer to **RQ2** and **RQ3**.

RQ2. Many NER tasks only require coarse grain entity extraction, which means that the labels used can be generic. Usually, commercial tools with pretrained models are trained to identify and extract the default set of entity types in NER, i.e., LOC, PER, MISC and ORG. However, for anonymization or information retrieval tasks, ad hoc models are required to identify tailored domain tags that are not available in commercial tools, as is the case in our datasets.

Previous contributions from the literature in languages other than English showed that the employment of pretrained models yields good accuracy (for an example in the Portuguese language, see [79]). In contrast to our scenario, these applications did not require NER types other than those offered by the pretrained model, which reduces overall accuracy. In fact, our experiments with pretrained models (which required mapping the police tags to the general tags) report mixed results: the F1 for the PER tag is 0.8796, is a remarkable result but is still worse than the best ad hoc model found; the F1 for the LOC tag is a mediocre 0.4479; Finally, the anonymization coverage only reaches 0.66, which is not acceptable for any practical purpose.

In conclusion, according to our results, the answer to **RQ2** is that it is necessary to train ad hoc models.

RQ3. The main results on ad hoc models are summarized in the following:

- Regarding the Tensorflow implemented models, our results are very similar to those of Pérez-Díez et al. [42], as our best F1 in the police domain is 0.915 (TENS3) while theirs is 0.9263. Even though these results are remarkable, they are lower than the best reported FlairNLP model, also, the latter is preferred as it is easier to integrate it in the tool which makes easier to reuse the model once trained, also it is lighter and faster.
- Regarding the FlairNLP framework embedding based configuration, our best configuration is FLAI4, which yields an F1 of 0.9294 and an anonymization coverage of 0.98962 in the police domain. As previously stated, this architecture is quite novel, as the state of the art usually defaults to using LSTM cells (as seen in Section 2) and only Ahmed et al. [40] reported an improvement with GRU cells.
- In terms of the Huggingface's Transformers, model TRAN2 achieves an F1 of 0.90269. This outcome is far better than the work by Ahmed et al. [46], the only other NER trans former-based technique. However, the transformer approach performs poorer than our earlier tested approaches, but it provides almost perfect anonymization coverage with only a 1% leak.

Given the previous insights, the best ad hoc model is a BiGRU_CRF with Flair Embeddings (FLAI4) as it achieves the best F1 and appears to be as robust as the LSTM models in the state of the art. Therefore, this model is deployed in the AGORA tool for the police domain (see Section 6).

5.2. Medical domain

The results of the experiments on the *MEDDOCAN Dataset* allow us to answer research question **RQ4**. Concretely, we studied:

- The performance of the previously obtained best model in the new domain.
- Whether the best architecture and configurations are transferable across domains.
- The performance of our best model in the medical domain by participating in the MEDDOCAN contest [78].

Table 18
Performance of FLAI4 trained on the police domain and applied to the MEDDOCAN task.

(a) Complete coverage.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.5613	0.3847	0.4909	0.5310
STREET_NAME	0.0437	0.0266	0.0331	
GENERAL_LOCATION	0.7524	0.3313	0.4600	
PERSON_NAME	0.8879	0.7961	0.8395	

(b) Partial coverage.

	Precision	Recall	F1	Anonymization coverage
AVERAGE SCORES	0.9177	0.6123	0.7240	0.7311
STREET_NAME	0.9286	0.5666	0.7038	
GENERAL_LOCATION	0.9232	0.4624	0.6162	
PERSON_NAME	0.9012	0.8081	0.8521	

Therefore, knowing that our best model is a BiGRU_CRF with Flair embeddings (FLAI4) we tested this pretrained model on the MEDDOCAN dataset. Our objective with this experiment was to see how the model trained with the police corpus behaves on other datasets. With this, we checked the robustness of the model when designing our tool, testing if it is necessary to train ad hoc models for each type of domain on which it is going to be used.

Given that the tags are not the same, we performed a mapping of the police domain tags to their counterparts in the *MEDDOCAN Dataset*:

- (ASSISTANCE SUBJECT'S NAME, ASSISTANCE SUBJECT'S RELATIVES, SANITARY PERSONEL'S NAMES)$\mapsto$ PERSON_NAME
- (HOSPITAL, TERRITORY, COUNTRY) $\mapsto$ GENERAL_LOCATION
- STREET_NAME $\mapsto$ STREET

For this analysis, the dataset is split in 500/250/250 for training/testing/validation (AKA development, in the context of the MEDDOCAN competition) purposes. The total performance is assessed using the F1 score.

Table 18(a) and 18(b) show the results on the *MEDDOCAN Dataset* of the FLAI4 model trained on the police domain dataset, computed using complete and partial coverage, respectively. Aside from the obvious issue that a model that is not specifically designed for this task does not cover all of the tags, the complete coverage performance (Table 18(a)) on the tags that are available is quite poor in comparison to other works on this corpus [78]. The same behavior is also observed when generic models are applied to the police domain, as seen on Section 5.1.1.

The partial coverage score (Table 18(b)) of "STREET_NAME" is much better than the complete one. This result shows that the model is capable of obtaining most of the entities. The performance difference is due to a lack of consistency in the format of "STREET_NAME" (e.g., different symbols and formats used to identify street numbers, apartment numbers).

5.2.1. Discussion

These results, are far from what we achieved using ad hoc models on the police domain. Hence, the answer to **RQ4** is that cross-domain models could be used minimally for anonymization if there is not a better alternative available. However, when having a corpus to train, it is better for the anonymization tool to incorporate an interchangeable model module for each optional domain.

5.2.2. MEDDOCAN contest

We trained all the considered model configurations (see Sections 3.2, 3.3 and 3.4) on the MEDDOCAN dataset. The models giving the best result have been enrolled in the MEDDOCAN contest [78] and evaluated using the CODALAB Evaluation Script.[11] It is important to notice that, to improve the performance of the model, we have performed corpus pre-processing and customized the tokenization functions of FlairNLP.

Our best model is FLAI1, therefore, this model is deployed in the AGORA tool for the medical domain. The model, achieved in the original contest [78] a precision, recall and F1 of 0.9495, 0.9399 and 0.9447 in the NER subtrack, coming in sixth position out of a total of 19 competitors. The same model earns a precision, recall and F1 of 0.9556, 0.9495 and 0.9507 in the SPANS sub-track, where we are ranked sixth overall in the contest.

6. AGORA

This section provides an overview of our created tool's strengths resulting from the constraints discovered in the state of the art analysis (Section 2), as well as a brief description of the features contained in it. The tool's demo video is available at https://www.youtube.com/watch?v=gHfA840WRd4, where readers can learn about and visualize AGORA's main workflow.

6.1. Comparison with state of the art tools

To the best of our knowledge, before AGORA, the state of the art did not offer a tool that integrated a workflow that allowed to anonymize, extract, and visualize named entities. Regarding geoparsing, [80] and [81] are the only projects that offer a visualization of the geoparsed and geocoded streets, using OpenStreetMaps. However, their tools do not provide all the features included in AGORA. Regarding anonymization, AGORA overcomes the following shortcomings of existing initiatives: First, they do not combine this task with information extraction. Furthermore, they do not provide public code, making these tools irreplicable. Additionally, they do not have filters that the user can select or change in order to adapt and clean the extracted information and then easily save it into files. Furthermore, the extracted locations are usually coarse grain (can identify areas, e.g. cities and regions) and cannot detect specific locations (i.e., addresses) and points of interest. Also, visualization tools are not available for the extracted data. In conclusion, there has been no attempt so far to combine all the above features into a single platform. Finally, since the majority of work in the anonymization domain is done on confidential and sensitive reports, most models are not public and freely available.

6.2. AGORA's features

The key characteristics of AGORA are next detailed:

Domain selection: AGORA enables the user to work across multiple domains. Its primary structure and operation are domain-agnostic. However, according to the findings of this work, for optimal performance, ad hoc NER models trained particularly for each domain should be used. We have so far implemented models in the police and medical domains. As a result, the user is initially requested to choose a domain from among the available models.

Document selection: AGORA allows users to work on one or more documents at the same time after selecting them from a folder on their computer.

Information extraction: AGORA performs NER on the user's selected batch of documents using the best model for the chosen domain. For this task, the system allows the user to select the entity types to be retrieved through a checkbox filter based on those that have been specifically tagged for the selected domain.

Document anonymization: AGORA performs NER on the user's selected batch of documents using the best model for the chosen domain. Then, the system allows the user to select the entity types to be anonymized through a checkbox filter based on those that have been specifically tagged for the selected domain.

Information export: AGORA allows users to export both the anony mized documents, which are produced as Word files, and the extracted tabular data (entities), which are exported as Excel files that display the entity, its label and its location in the text.

Information visualization: AGORA makes use of the Google Maps API to display the geographical entities retrieved from each document on an interactive map.

Information enrichment: AGORA allows for additional information to be visualized, extracted and exported. In particular, the user can include and represent information regarding relevant places (e.g., in the police domain, the victim's address, or the last known whereabouts), as well as amenities in a specified radius (e.g., banks, restaurants, schools, parks). The latter are obtained from OpenStreetMaps API.

7. Conclusions

Among institutions, there is a need for data sharing that complies with data protection laws. This research addresses this issue by developing a cutting-edge tool that anonymizes sensitive documents, extracts key tabulated information, and displays extracted locations.

To accomplish this, the literature on the subject has been thoroughly reviewed. On top of helping us understanding that both document anonymization and entities geocoding rely on underlying NER models, it has allowed us to identify gaps that have been addressed in this research. In particular, we have conducted extensive computational experiments to compare NER model configurations previously presented in the literature, as well as our own proposed architectures. The models have been validated and tested in two different domains, i.e., police (new to NER) and medical (both datasets are in Spanish). The heterogeneity in the corpora in terms of structure, contents and authorship requires the development of a robust methodology based on neural networks. Our experiments point out that the most successful models are those trained ad hoc in the FlairNLP framework, using a Bidirectional LSTM or GRU network with a CRF classifier combined with advanced embeddings (Spanish FastText Word or Flair are the best ones in the revised domains). In the police domain, we achieve an F1 of 0.9294, rivaling state of the art approaches, while the anonymization coverage is almost perfect and only 1% of sensitive information is leaked (i.e. a human reviewer must review the document for achieving perfect anonymization). In the medical domain, we obtained an F1 of 0.9507; our model ranked 6th (out of 19) in the MEDDOCAN contest.

The main output of this research is the AGORA tool, an intelligent system developed in collaboration with ONDOD that allows processing of sensitive documents in the policing and medical domains. AGORA overcomes the limits of previous systems in the state of the art, as it integrates entity filtering, extraction, and visualization. The system relies on our best NER models (one per domain) to identify and extract entities. Additionally, the tool is capable of obtaining and visualizing amenities on a map. All the information can be exported for post-processing and analysis. AGORA can be easily extended to other domains by training a new NER model on a specific dataset.

[11] Available at https://github.com/PlanTL-GOB-ES/SPACCC_MEDDOCAN/tree/master/scripts/CODALAB-Evaluation-script, last accessed May 24, 2023.

We hope that this work will be a useful source of ideas for future research on document anonymization and NER, and will contribute further in the development of applied intelligent systems in the real world.

CRediT authorship contribution statement

Rodrigo Juez-Hernandez: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Lara Quijano-Sánchez:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Federico Liberatore:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Jesús Gómez:** Conceptualization, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Processing of documents containing personal data was done in accordance with the institutions' security protocols as well as the security measures stipulated by national and European laws. The research of Quijano-Sánchez was conducted with financial support from the Spanish Ministry of Science and Innovation, grant PID2019-108965GB-I00. The research of Liberatore is funded by Spanish Ministry of Science and Innovation, grant PID2019-108679RB-I00. All the financial support is gratefully acknowledged. The authors would like to thank the ONDOD and the Spanish National Police for all the help and resources provided to this research project.

References

[1] M. Xu, J.M. David, S.H. Kim, et al., The fourth industrial revolution: Opportunities and challenges, Int. J. Financ. Res. 9 (2) (2018) 90–95.
[2] H. Fröhlich, R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M.H. Maathuis, Y. Moreau, S.A. Murphy, T.M. Przytycka, et al., From hype to reality: Data science enabling personalized medicine, BMC Med. 16 (1) (2018) 1–15.
[3] S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M.N.K. Boulos, A. Weller, et al., Leveraging data science to combat COVID-19: A comprehensive review, IEEE Trans. Artif. Intell. 1 (1) (2020) 85–103.
[4] X. Li, D. Zhang, Y. Zheng, W. Hong, W. Wang, J. Xia, Z. Lv, Evolutionary computation-based machine learning for smart city high-dimensional big data analytics, Appl. Soft Comput. (2022) 109955.
[5] G.J. Smith, L. Bennett Moses, J. Chan, The challenges of doing criminology in the big data era: Towards a digital and data-driven approach, Br. J. Criminol. 57 (2) (2017) 259–274.
[6] G. Ridgeway, Policing in the era of big data, Annu. Rev. Criminol. 1 (2018) 401–419.
[7] L. Quijano-Sánchez, F. Liberatore, G. Rodríguez-Lorenzo, R.E. Lillo, J.L. González-Álvarez, A twist in intimate partner violence risk assessment tools: Gauging the contribution of exogenous and historical variables, Knowl.-Based Syst. 234 (2021) 107586.
[8] E. Lima, T. Vieira, E. de Barros Costa, Evaluating deep models for absenteeism prediction of public security agents, Appl. Soft Comput. 91 (2020) 106236.
[9] C. Garvie, J. Frankle, Facial-recognition software might have a racial bias problem, Atlantic 7 (2016).
[10] G. Zhang, X. Zhu, L. Yin, W. Pedrycz, Z. Li, Granular data representation under privacy protection: Tradeoff between data utility and privacy via information granularity, Appl. Soft Comput. 131 (2022) 109808.
[11] F. Liberatore, M. Camacho-Collados, L. Quijano-Sánchez, Equity in the police districting problem: Balancing territorial and racial fairness in patrolling operations, J. Quant. Criminol. (2021).
[12] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, J.E. Gilbert, A review of predictive policing from the perspective of fairness, Artif. Intell. Law 30 (1) (2022) 1–17.
[13] G.S. Nelson, Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification, in: SAS Global Forum Proceedings, 2015, pp. 1–23.
[14] M.M. Anjum, N. Mohammed, X. Jiang, De-identification of unstructured clinical texts from sequence to sequence perspective, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 2438–2440.
[15] C. Weng, C. Friedman, C.A. Rommel, J.F. Hurdle, A two-site survey of medical center personnel's willingness to share clinical data for research: Implications for reproducible health NLP research, BMC Med. Inform. Decis. Mak. 19 (3) (2019) 70.
[16] C. Kadar, R. Maculan, S. Feuerriegel, Public decision support for low population density areas: An imbalance-aware hyper-ensemble for spatio-temporal crime prediction, Decis. Support Syst. 119 (2019) 107–117.
[17] B. Mohit, Named entity recognition, in: Natural Language Processing of Semitic Languages, Springer, 2014, pp. 221–245.
[18] M. Karimzadeh, S. Pezanowski, A.M. MacEachren, J.O. Wallgrün, GeoTxt: A scalable geoparsing system for unstructured text geolocation, Trans. GIS 23 (2019) 118–136.
[19] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, Y. LeTraon, A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS, IEEE, 2019, pp. 338–343.
[20] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60.
[21] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, " O'Reilly Media, Inc.", 2009.
[22] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, 7 (1), 2017, pp. 411–420, in press.
[23] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: CoNLL, 2009.
[24] Apache Software Foundation, openNLP natural language processing library, 2014.
[25] B. Baldwin, K. Dayanidhi, Natural Language Processing with Java and LingPipe Cookbook, Packt Publishing, 2014.
[26] D.M.H. Cunningham, K. Bontcheva, Text Processing with GATE (Version 6), University of Sheffield D, 2011.
[27] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[28] T. Crayston, TextRazor, 2021, GitHub Repository, GitHub, https://www.textrazor.com/.
[29] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: NAACL Annual Conference, 2019, pp. 54–59.
[30] D. Samy, Reconocimiento y clasificación de entidades nombradas en textos legales en español, Procesamiento Del Lenguaje Natural 67 (2021) 103–114.
[31] V.A. Mozharova, N.V. Loukachevitch, Combining knowledge and CRF-based approach to named entity recognition in Russian, in: International Conference on Analysis of Images, Social Networks and Texts, Springer, 2016, pp. 185–195.
[32] M. Gridach, Deep learning approach for arabic named entity recognition, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 439–451.
[33] I.S. Azarine, M.A. Bijaksana, I. Asror, Named entity recognition on Indonesian tweets using hidden Markov model, in: 2019 7th International Conference on Information and Communication Technology, ICoICT, IEEE, 2019, pp. 1–5.
[34] E. Trandafili, E.K. Meçe, E. Duka, A named entity recognition approach for albanian using deep learning, in: Complex Pattern Mining, Springer, 2020, pp. 85–101.
[35] J. Santoso, E.I. Setiawan, C.N. Purwanto, E.M. Yuniarno, M. Hariadi, M.H. Purnomo, Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory, Expert Syst. Appl. 176 (2021) 114856.
[36] Z. Chen, B. Pokharel, B. Li, S. Lim, Location extraction from Twitter messages using a bidirectional long short-term memory neural network with conditional random field model, in: International Conference on Geographical Information Systems Theory, Applications and Management, Springer, 2020, pp. 18–30.

[37] C. Napoli, E. Tramontana, G. Verga, Extracting location names from unstructured italian texts using grammar rules and MapReduce, in: International Conference on Information and Software Technologies, Springer, 2016, pp. 593–601.
[38] T. Chomutare, Clinical notes de-identification: Scoping recent benchmarks for n2c2 datasets, Stud. Health Technol. Inf. 289 (2022) 293–296.
[39] Ö. Uzuner, A. Stubbs, Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks, J. Biomed. Inform. 58 Suppl (2015) S1–S5.
[40] T. Ahmed, M.M.A. Aziz, N. Mohammed, De-identification of electronic health record using neural network, Sci. Rep. 10 (1) (2020) 1–11.
[41] 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.
[42] 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.
[43] R. Catelli, F. Gargiulo, E. Damiano, M. Esposito, G. De Pietro, Clinical de-identification using sub-document analysis and ELECTRA, in: 2021 IEEE International Conference on Digital Health, IEEE, 2021, pp. 266–275.
[44] R. Catelli, V. Casola, G. De Pietro, H. Fujita, M. Esposito, Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification, Knowl.-Based Syst. 213 (2021).
[45] J. Santos, H. dos Santos, F. Tabalipa, R. Vieira, De-identification of clinical notes using contextualized language models and a token classifier, Lecture Notes in Comput. Sci. 13074 LNAI (2021) 33–41.
[46] A. Ahmed, A. Abbasi, C. Eickhoff, Benchmarking modern named entity recognition techniques for free-text health record deidentification, AMIA Summits Transl. Sci. Proc. 2021 (2021) 102.
[47] O. Sotolář, J. Plhák, D. Šmahel, Towards personal data anonymization for social messaging, in: International Conference on Text, Speech, and Dialogue, Springer, 2021, pp. 281–292.
[48] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, A novel COVID-19 data set and an effective deep learning approach for the de-identification of Italian medical records, IEEE Access (2021).
[49] F. Hassan, M. Jabreel, N. Maaroof, D. Sánchez, J. Domingo-Ferrer, A. Moreno, ReCRF: Spanish medical document anonymization using automatically-crafted rules and CRF, in: IberLEF@ SEPLN, 2019, pp. 727–734.
[50] 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.
[51] V. Foufi, C. Gaudet-Blavignac, R. Chevrier, C. Lovis, De-identification of medical narrative data, Stud. Health Technol. Inf. 244 (2017) 23–27.
[52] P. Richter-Pechanski, S. Riezler, C. Dieterich, De-identification of German medical admission notes, Stud. Health Technol. Inf. 253 (2018) 165–169.
[53] P. Richter-Pechanski, A. Amr, H. Katus, C. Dieterich, Deep learning approaches outperform conventional strategies in de-identification of German medical reports, Stud. Health Technol. Inf. 267 (2019) 101–109.
[54] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set, Appl. Soft Comput. 97 (2020) 106779.
[55] X. Yang, T. Lyu, Q. Li, C.-Y. Lee, J. Bian, W.R. Hogan, Y. Wu, A study of deep learning methods for de-identification of clinical notes in cross-institute settings, BMC Med. Inf. Decis. Mak. 19 (5) (2019) 232.
[56] F. Hassan, D. Sánchez, J. Soria-Comas, J. Domingo-Ferrer, Automatic anonymization of textual documents: Detecting sensitive information via word embeddings, in: IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2019, pp. 358–365.
[57] F. Dernoncourt, J.Y. Lee, O. Uzuner, P. Szolovits, De-identification of patient notes with recurrent neural networks, J. Am. Med. Inf. Assoc. 24 (3) (2016) 596–606.
[58] E. Eder, U. Krieg-Holz, U. Hahn, CodE Alltag 2.0—a pseudonymized German-language email corpus, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4466–4477.
[59] S. Hina, R. Asif, S.A. Ali, Anonymization framework for securing protected health information in a complex dataset of medical narratives, Mehran Univ. J. Eng. Technol. 39 (3) (2020) 612–624.
[60] C.A. Libbi, J. Trienes, D. Trieschnigg, C. Seifert, Generating synthetic training data for supervised de-identification of electronic health records, Future Internet 13 (5) (2021) 136.
[61] C. Liu, J. Li, Y. Liu, J. Du, B. Tang, R. Xu, Named entity recognition in clinical text based on capsule-LSTM for privacy protection, in: International Conference on AI and Mobile Services, Springer, 2019, pp. 166–178.
[62] I. Calapodescu, D. Rozier, S. Artemova, J.-L. Bosson, Semi-automatic de-identification of hospital discharge summaries with natural language processing: A case-study of performance and real-world usability, in: 2017 IEEE International Conference on Internet of Things, 2017, pp. 1106–1111.
[63] K. Lai, J.R. Porter, M. Amodeo, D. Miller, M. Marston, S. Armal, A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures, Inf. Process. Manage. 59 (1) (2022) 102735.
[64] L. Cadorel, A. Blanchi, A.G. Tettamanzi, Geospatial knowledge in housing advertisements: Capturing and extracting spatial information from text, in: Proceedings of the Knowledge Capture Conference, 2021, pp. 41–48.
[65] A. Molina-Villegas, V. Muñiz-Sanchez, J. Arreola-Trapala, F. Alcántara, Geographic named entity recognition and disambiguation in Mexican news using word embeddings, Expert Syst. Appl. 176 (2021) 114855.
[66] L. Moncla, M. Gaio, T. Joliveau, Y.-F.L. Lay, Automated geoparsing of paris street names in 19th century novels, in: Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, 2017, pp. 1–8.
[67] S.E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, Y. Kompatsiaris, Location extraction from social media: Geoparsing, location disambiguation, and geotagging, ACM Trans. Inf. Syst. 36 (4) (2018).
[68] E. Aldana-Bobadilla, A. Molina-Villegas, I. Lopez-Arevalo, S. Reyes-Palacios, V. Muñiz-Sanchez, J. Arreola-Trapala, Adaptive geoparsing method for toponym recognition and resolution in unstructured text, Remote Sens. 12 (18) (2020) 3041.
[69] R. Schwarzenberg, L. Hennig, H. Hemsen, In-memory distributed training of linear-chain conditional random fields with an application to fine-grained named entity recognition, in: International Conference of the German Society for Computational Linguistics and Language Technology, Springer, 2017, pp. 155–167.
[70] D. Cabo, GitHub thread evaluating the performance of stock NER tools available, 2022, GitHub Repository, GitHub, https://github.com/civio/verba/issues/10.
[71] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, 2015, arXiv.
[72] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, 2016.
[73] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, 2016.
[74] H. Patel, BioNerFlair: Biomedical named entity recognition using flair embedding and sequence tagger, 2020, arXiv.
[75] 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.
[76] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC At ICLR 2020, 2020.
[77] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, CoRR abs/1810.04805.
[78] 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.
[79] J. de Oliveira Lima, C. da Silveira Colombo, F. Izo, J.C.P. Pirovani, E. de Oliveira, Using CRF+LG for automated classification of named entities in newspaper texts, in: 2020 XLVI Latin American Computing Conference, CLEI, IEEE, 2020, pp. 27–32.
[80] A. Girsang, S. Isa, R. Fajar, Implementation of a geocoding in journalist social media monitoring system, Int. J. Eng. Trends Technol. 69 (12) (2021) 103–113.
[81] B. Alex, C. Grover, R. Tobin, J. Oberlander, Geoparsing historical and contemporary literary text set in the City of Edinburgh, Lang. Resour. Eval. 53 (4) (2019) 651–675.