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 Autónoma 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

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 novelty proposed ones. Finally, the methodology, conclusions and software provided can be easily reused in similar application scenarios.

© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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...
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 anonymization, geoparsing, geocoding and visualization and is domain-agnostic.
2. An extensive analytical literature review on NER, anonymization, 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 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://huggingface.co/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].

<table>
<thead>
<tr>
<th>NER Tool</th>
<th>Architecture</th>
<th>Corpus for base version</th>
<th>Other languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford NLP</td>
<td>CRF</td>
<td>CoNLL2003, Ritter, MSM2013</td>
<td>German, Spanish, Chinese</td>
</tr>
<tr>
<td>NLTK</td>
<td>Max Entropy Model</td>
<td>Stanford NER</td>
<td>German, Spanish, Chinese</td>
</tr>
<tr>
<td>SpaCy [22]</td>
<td>CNNs (2.0), Transformers (3.0)</td>
<td>Wikigold, OnToNotes</td>
<td>Catalan, Chinese, Dutch, Finnish, French, German, Greek, Italian, Japanese, Korean, Lithuanian, Macedonian, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish</td>
</tr>
<tr>
<td>CogComp [23]</td>
<td>Perceptron</td>
<td>CoNLL2003, Enron email</td>
<td>NO</td>
</tr>
<tr>
<td>OpenNLP [24]</td>
<td>Maximum Entropy Model</td>
<td>CoNLL2003</td>
<td>Danish, German, Spanish, Portuguese, Swedish, Dutch</td>
</tr>
<tr>
<td>LingPipe [25]</td>
<td>Hidden Markov Models</td>
<td>MUC-6</td>
<td>Arabic, Chinese, Dutch, German, Greek, Hindi, Japanese, Korean, Portuguese and Spanish</td>
</tr>
<tr>
<td>GATE [26]</td>
<td>Rule Based</td>
<td>N/A</td>
<td>NO</td>
</tr>
<tr>
<td>TextRazor [28]</td>
<td>N/A</td>
<td>MSM2013, Ritter, UMBBC</td>
<td>Arabic, German, French, Spanish, Dutch, Danish, Malayalam, Portuguese</td>
</tr>
<tr>
<td>Flair [29]</td>
<td>BiLSTM, Transformers</td>
<td>CoNLL2003, OnToNotes</td>
<td></td>
</tr>
</tbody>
</table>

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

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

#### 2.1.3. NER-based tools
Given that AGORA aims to enhance multidisciplinary collaborations and provide an 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

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

Table 4

<table>
<thead>
<tr>
<th>Paper</th>
<th>Language</th>
<th>Domain</th>
<th>Best model</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>[63]</td>
<td>English</td>
<td>News</td>
<td>Hand Crafted Rules + Domain Specific ML Flair+CamemBERT+Word2Vec + BiLSTM+CRF</td>
<td>F1 0.7213, Accuracy 0.73</td>
</tr>
<tr>
<td>[64]</td>
<td>French</td>
<td>Housing Advertisements.</td>
<td>ConEc Embeddings + 1 hidden Layer Neural Network “PERDIDO” platform</td>
<td>Precision 0.863, Recall 0.889, F10.876</td>
</tr>
<tr>
<td>[65]</td>
<td>Spanish</td>
<td>News</td>
<td>map–database entity matching (OpenStreetMaps gazetteer match)</td>
<td>Accuracy 0.9633, Precision 0.893</td>
</tr>
<tr>
<td>[66]</td>
<td>French</td>
<td>19th century Novels.</td>
<td>Perceptron with 3 hidden units Linear-chain CRF</td>
<td>F1 0.5983, Recall 0.5663</td>
</tr>
<tr>
<td>[67]</td>
<td>English/Turkish</td>
<td>Social Media</td>
<td>map–database entity matching (OpenStreetMaps gazetteer match)</td>
<td>Precision 0.96–0.99, F1 0.90–0.97</td>
</tr>
<tr>
<td>[68]</td>
<td>Spanish</td>
<td>Geography</td>
<td>Perceptron with 3 hidden units Linear-chain CRF</td>
<td>F1 0.5983, Recall 0.5663</td>
</tr>
<tr>
<td>[69]</td>
<td>German</td>
<td>Tweets</td>
<td>Perceptron with 3 hidden units Linear-chain CRF</td>
<td>Recall 0.504, Standford NER precision 0.79</td>
</tr>
</tbody>
</table>

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 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] Lampel 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. 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.

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 corporuses 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.

---

\(^6\) 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).

\(^8\) https://huggingface.co/docs/transformers/index

\(^9\) This self-fine tuned BETO model for NER can be found at [huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner](https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner).
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 allow us to answer research questions RQ2 and RQ3. For this analysis, the dataset is split in 250/50/19 for training/testing/validation 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.

The performance scores have been calculated using the tailormade “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) ➔ LOC
- PERSON_NAME ➔ 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 Spanish NER model Flair outperforms the Stanza (StanfordNLP) 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 Stanford NER model to an ad hoc trained model, with the latter achieving significantly better results.

### 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].

---

**Table 5**

Test performance results for pretrained model Flair.

<table>
<thead>
<tr>
<th>(a) Complete coverage</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE SCORES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>0.4523</td>
<td>0.5420</td>
<td>0.4270</td>
<td></td>
</tr>
<tr>
<td>PER</td>
<td>0.7821</td>
<td>0.9310</td>
<td>0.8501</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6**

Test performance results for pretrained model Stanza.

<table>
<thead>
<tr>
<th>(a) Complete coverage</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE SCORES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>0.4908</td>
<td>0.65</td>
<td>0.5593</td>
<td>0.6747</td>
</tr>
<tr>
<td>PER</td>
<td>0.3316</td>
<td>0.4464</td>
<td>0.3805</td>
<td></td>
</tr>
</tbody>
</table>

**Table 7**


<table>
<thead>
<tr>
<th>STREET_NAME</th>
<th>GENERAL_LOCATION</th>
<th>PERSON_NAME</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7722</td>
<td>0.6466</td>
<td>0.8071</td>
<td>0.7712</td>
<td>0.6643</td>
<td>0.7142</td>
<td>0.8695</td>
</tr>
</tbody>
</table>

**Table 8**

Test scores of best model for TENS2. Best hyperparameter configuration: (LSTM × CHAR size) = (100 × 100).

<table>
<thead>
<tr>
<th>STREET_NAME</th>
<th>GENERAL_LOCATION</th>
<th>PERSON_NAME</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9163</td>
<td>0.8595</td>
<td>0.9447</td>
<td>0.9128</td>
<td>0.9092</td>
<td>0.9128</td>
<td>0.9743</td>
</tr>
</tbody>
</table>

---

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.
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 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 the other configurations as seen when comparing the LSTM models (i.e., FLAI1, FLAI2, and FLAI3).

### Tables

#### Table 9
<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.9251</td>
<td>0.9051</td>
<td>0.915</td>
</tr>
<tr>
<td>STREET_NAME</td>
<td>0.8247</td>
<td>0.8554</td>
<td>0.8398</td>
</tr>
<tr>
<td>GENERAL LOCATION</td>
<td>0.9120</td>
<td>0.8704</td>
<td>0.8907</td>
</tr>
<tr>
<td>PERSON_NAME</td>
<td>0.9589</td>
<td>0.9401</td>
<td>0.9494</td>
</tr>
</tbody>
</table>

#### Table 10
<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.9266</td>
<td>0.9274</td>
<td>0.9072</td>
</tr>
<tr>
<td>STREET_NAME</td>
<td>0.9174</td>
<td>0.9136</td>
<td>0.9155</td>
</tr>
<tr>
<td>GENERAL LOCATION</td>
<td>0.8955</td>
<td>0.8969</td>
<td>0.8962</td>
</tr>
<tr>
<td>PERSON_NAME</td>
<td>0.967</td>
<td>0.9717</td>
<td>0.9693</td>
</tr>
</tbody>
</table>

#### Table 11
<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.9086</td>
<td>0.906</td>
<td>0.9072</td>
</tr>
<tr>
<td>STREET_NAME</td>
<td>0.8816</td>
<td>0.8926</td>
<td>0.8871</td>
</tr>
<tr>
<td>GENERAL LOCATION</td>
<td>0.8925</td>
<td>0.8821</td>
<td>0.8873</td>
</tr>
<tr>
<td>PERSON_NAME</td>
<td>0.9516</td>
<td>0.9432</td>
<td>0.9474</td>
</tr>
</tbody>
</table>

#### Table 12
<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.9173</td>
<td>0.9266</td>
<td>0.9219</td>
</tr>
<tr>
<td>STREET_NAME</td>
<td>0.9184</td>
<td>0.9259</td>
<td>0.9221</td>
</tr>
<tr>
<td>GENERAL LOCATION</td>
<td>0.8779</td>
<td>0.8881</td>
<td>0.8829</td>
</tr>
<tr>
<td>PERSON_NAME</td>
<td>0.9556</td>
<td>0.9659</td>
<td>0.9607</td>
</tr>
</tbody>
</table>

#### Table 13
<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Anonymization coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>0.9251</td>
<td>0.9338</td>
<td>0.9294</td>
</tr>
<tr>
<td>STREET_NAME</td>
<td>0.8988</td>
<td>0.9136</td>
<td>0.9061</td>
</tr>
<tr>
<td>GENERAL LOCATION</td>
<td>0.9074</td>
<td>0.9113</td>
<td>0.9094</td>
</tr>
<tr>
<td>PERSON_NAME</td>
<td>0.969</td>
<td>0.9766</td>
<td>0.9728</td>
</tr>
</tbody>
</table>

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.

#### 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 Poli ce 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 implementation 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 transformer-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.

<table>
<thead>
<tr>
<th><strong>Table 18</strong> Performance of FLAI4 trained on the police domain and applied to the MEDDOCAN task.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(a) Complete coverage.</strong></td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>AVERAGE SCORES</td>
</tr>
<tr>
<td>STREET_NAME</td>
</tr>
<tr>
<td>GENERAL_LOCATION</td>
</tr>
<tr>
<td>PERSON_NAME</td>
</tr>
<tr>
<td><strong>(b) Partial coverage.</strong></td>
</tr>
<tr>
<td>Precision</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>AVERAGE SCORES</td>
</tr>
<tr>
<td>STREET_NAME</td>
</tr>
<tr>
<td>GENERAL_LOCATION</td>
</tr>
<tr>
<td>PERSON_NAME</td>
</tr>
</tbody>
</table>

- 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].

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 domain tags' counterparts in the MEDDOCAN Dataset:

- (ASSISTANCE SUBJECT’S NAME, ASSISTANCE SUBJECT’S RELATIVES, SANITARY PERSON’S NAMES) → PERSON_NAME
- (HOSPITAL, TERRITORY, COUNTRY) → GENERAL_LOCATION
- STREET_NAME → 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. 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 FairNLP.

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=HFAB40WRd14, 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 irreproducible. Additionally, they do not have filters that the user can select or change in order to adapt and clean the extracted information. Finally, 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 anonymized 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.

---

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
