Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3751–3760
Marseille, 20-25 June 2022

Spanish Datasets for Sensitive Entity Detection in the Legal Domain

Ona de Gibert[1], **Aitor García-Pablos**[2], **Montse Cuadros**[2], **Maite Melero**[1]
[1]Barcelona Super Computing Center (BSC),
[2]SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Plaça Eusebi Güell 1-3, Barcelona 08034, Spain
Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain
ona.degibert@bsc.es, agarciap@vicomtech.org, mcuadros@vicomtech.org, maite.melero@bsc.es

Abstract

The de-identification of sensible data, also known as automatic textual anonymisation, is essential for data sharing and reuse, both for research and commercial purposes. The first step for data anonymisation is the detection of sensible entities. In this work, we present four new datasets for named entity detection in Spanish in the legal domain. These datasets have been generated in the framework of the MAPA project, three smaller datasets have been manually annotated and one large dataset has been automatically annotated, with an estimated error rate of around 14%. In order to assess the quality of the generated datasets, we have used them to fine-tune a battery of entity-detection models, using as foundation different pre-trained language models: one multilingual, two general-domain monolingual and one in-domain monolingual. We compare the results obtained, which validate the datasets as a valuable resource to fine-tune models for the task of named entity detection. We further explore the proposed methodology by applying it to a real use case scenario.

Keywords: Anonymisation, Named Entity Recognition, Sensitive Data Detection

1. Introduction

The de-identification of sensible data, also known as automatic textual anonymisation, is essential for data sharing and reuse both for research and commercial purposes, in order to comply with the General Data Protection Regulation (GDPR) law[1], which protects data privacy. Anonymisation helps tackle the issue of protecting personal information, by identifying and subsequently processing (for example, by ofuscating) personal data such as names, emails, addresses, etc. from a text, while preserving its formal structure.

In the context of the European Union, where Public Administrations are encouraged to re-use and open public sector information[2], there is a need to find effective ways of sharing large amounts of collected information, while complying with the GDPR. The project MAPA (Multilingual Anonymisation for Public Administrations)[3], an INEA-funded Action for the European Commission under the Connecting Europe Facility (CEF) program, aims at addressing this issue by targeting domain-specific anonymisation in the 24 languages of the European Union (Gianola et al., 2020).

The first step for data anonymisation is the detection of sensible entities that need to be anonymised. In this work, we present several datasets annotated for named entity detection in Spanish, with a view to train models for anonymisation of personal information in the legal domain. These datasets have been developed in the framework of the MAPA project and are available to the community. We have used them to fine-tune different Transformer-based pre-trained language models, resulting in a set of entity-detection models, which we have evaluated. We compare the results obtained in our experiments, to showcase the impact of the language and in-domain variations among the pre-trained language models and the datasets.

The remaining of this article is organised as follows. In Section 2, we summarize the relevant related work. In Section 3, we describe in detail the datasets collected. Section 4 focuses on the annotation method and in Section 5 we go through the experiments and the results of the evaluation. Section 6 describes the application of our method to a real use case. Finally, in sections 7 and 8, we discuss the obtained results and propose future work.

2. Related Work

Anonymisation refers to the task of identifying and processing sensitive data within a given text. Medlock (2006) proposes three different de-identification approaches when dealing with sensible or private information, as illustrated in Table 1.

On the one hand, anonymisation removal refers to the process by which personal information is simply removed from a text, as a way to protect personal information (Ji et al., 2017). Similarly, categorical anonymisation refers to the procedure by which entities in a text are replaced with categories identifying the type of private data, such as PERSON, DATE or ORGANISATION, for example. Finally, in pseudo-anonymisation, identifiable data is not only removed, but also replaced by new artificial entities (Francopoulo and Schaub, 2020). Pseudo-anonymisation allows to reuse the data for Natural Language Processing (NLP), since the result is a natural text and the re-identification of the original person is no longer possible, which means the data is available for sharing while complying with GDPR.

[1] Council Regulation 2016/679, 2016 O.J. (L 119) (EU) 1

[2] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information

[3] Grant Agreement No INEA/CEF/ICT/A2019/1927065.

Original	*El pasado día 3 de junio de 2008, en la zona de discotecas de Sant Andreu, [...] agredieron a José Antonio Sánchez Pujol.*
Removal	*El pasado día , en la zona de discotecas de , agredieron a .*
Categorical	*El pasado día DATE, en la zona de discotecas de ADDRESS, [...] agredieron a PERSON.*
Pseudo-anonymisation	*El pasado día 24 de marzo de 2014, en la zona de discotecas de Port Olímpic, [...] agredieron a Aaron Leon Rodriguez.*

Table 1: Illustrated de-identification approaches proposed by Medlock (2006).

Anonymisation is typically approached as a Named Entity Recognition and Classification (NERC) task, which is a well-known NLP task. Most works about anonymisation in the literature focus on the medical domain (e.g. (Lima-López et al., 2020; Marimon et al., 2019)), because of its sensible nature. Domain-specific anonymisation still presents challenges, due to the general lack of resources for training and evaluation. A common approach to address domain-adapted NERC is to use a pre-trained language model, such as BERT (Devlin et al., 2019), and fine-tune it with domain-specific data (Mao and Liu, 2019).

In this work, we focus on NERC within the legal domain. In this domain, Chalkidis et al. (2020) developed a set of legal BERTs for English, by pre-training the language model on in-domain corpora; a similar approach was proposed by Bambroo and Awasthi (2021), by further fine-tuning DistilBERT (Sanh et al., 2019).

When evaluating NERC in Spanish, the data of the CoNLL-2002 Shared Task (Tjong Kim Sang, 2002) is often used. In recent years, there has been an increased effort to develop resources for evaluation of downstream tasks in Spanish. To the best of our knowledge, currently there only exists one public resource in the legal domain for Spanish, Legal-ES (Samy et al., 2020), which is a set of resources for Spanish legal text processing, including a large corpus of over 2000 million words, word embeddings and topic models. Recently, Gutiérrez-Fandiño et al. (2021) have developed a Spanish legalese language model, RoBERTaLex, trained on legal-domain corpora from different sources.

Nevertheless, to our knowledge, there exist no open resources annotated for NERC in Spanish in the legal domain. With the present contribution, we intend to fill this gap. With the release of the created resources for fine-tuning and evaluation of sensitive entities detection in the legal domain, we expect to encourage the development of domain-adapted anonymisation tools for Spanish in this field.

3. Datasets

In this work, we present four different datasets annotated to train models for detection of sensitive entities in Spanish in the legal domain, and eventually anonymise them.

We present three new manually annotated datasets plus an automatically annotated dataset for NERC, following a double hierarchy entity distribution proposed within the MAPA project (Gianola et al., 2020). In section 6, we describe a real use case scenario that confirms the applicability of these types of resources.

Corpus de Procesos Penales (CPP)
This is a Spanish corpus of 10 court cases (Taranilla, 2012). Each case consists of five parts: the public prosecutor, the private prosecution, the defense, the oral trial and the final sentence and the verdict. We used all but the oral trial section for annotation. To make the documents suitable for annotation, we have preprocessed the original PDF files by extracting plain text format with the library `pdftotext`.[4] After recovering broken sentences, we have applied sentence splitting. The total number of documents is 35, accounting for 853 sentences.

Spanish annotated subset of EUR-Lex corpus
EUR-Lex (Baisa et al., 2016) is a multilingual corpus of court decisions and legal dispositions in the 24 official languages of the European Union. We have manually annotated a subset of 12 documents in Spanish, totalling 2,261 sentences, following the guidelines of the MAPA project. The selection of the documents was linguistically motivated: since MAPA is a multilingual project, the documents that existed in all the languages of the European Union were chosen.

Dictámenes del Estado (DE-Silver)
This is a pseudo-anonymized corpus of Opinions from the State Council, extracted from Samy et al. (2020), containing 71,667 documents (or 2,711,885 sentences). The original corpus, before annotation, contains manually annotated placeholders for personal names, schools, streets and companies. In order to be able to use this corpus, we have inserted fabricated names in lieu of the placeholders using a list of synthetic person names. To build the list of synthetic names we have used three gazzeeteers provided by the Spanish National Institute of Statistics[5]: a list of male names, a list of female names and a list of surnames. The final list is composed of full names consisting of a

[4] `https://pypi.org/project/pdftotext/`

[5] `ww.ine.es/`

	CPP		EUR-Lex		DE-Silver-5k/ DE-Gold	
	Train	Test	Train	Test	Train	Test
Docs	32	3	9	3	5,000	10
Tokens	40,573	7,445	62,705	27,454	10,943,332	16,592
Level 1 Tags	3,189	578	2,837	1,306	907,920	1,367
Level 2 Tags	2,463	486	1,721	910	604,598	1,039
Level 1 Ents	921	191	1,076	493	266,336	388
Level 2 Ents	1,555	338	1,495	826	471,335	716

Table 2: Statistics of the datasets used for our experiments. 'Tags' refers to the individual tokens that are annotated, while 'Ents' refers to whole entities.

first name and two surnames, following the usual convention in Spain. In order to be more realistic we have randomly chosen names and surnames, but following frequency criteria and Zipf's Law. When replacing a person name, we have used the full name with two surnames, when replacing schools and company names we have used name + surname. At all times we aim for maintaining gender coherence.

The total amount of tokens of the resulting corpus sums up to 135,348 million tokens, which makes the corpus suitable for pre-training language models (Gutiérrez-Fandiño et al., 2021).

Annotated subset of Dictámenes del Estado (DE-Gold)

Once the automatic insertion of entities as described above was completed, we manually annotated 10 documents (306 sentences) to use as a Gold Standard (DE-Gold). We use a subset of 5,000 documents from DE-Silver (DE-Silver-5k) as training set and DE-Gold as test set.

The statistics for all the datasets, after splitting into train and test, are shown in table 2.

4. Annotation

The annotation scheme consists of a complex two level hierarchy adapted to the legal domain, it follows the scheme described in (Gianola et al., 2020), illustrated in Figure 3.

Level 1 entities refer to general categories (PERSON, DATE, TIME, ADDRESS...) and level 2 entities refer to more fine-grained subcategories (given name, personal name, day, year, month...). Eur-Lex, CPP and DE have been annotated following this annotation scheme, thus they can be used either separately or merged together, depending on the needs of the model to be trained.

The entity distribution per dataset can be found in Table 3, for level 1 and level 2 entities.

The manual annotation was performed using INCeP-TION (Klie et al., 2018) by a sole annotator following the guidelines provided by the MAPA consortium. An example of the INCEpTION annotation interface can be found in Figure 1.

4.1. Silver Standard

We also release a silver standard (DE-Silver) with documents of the DE automatically annotated with our best performing model for the DE-Gold dataset. Although we plan to publish the complete dataset containing over 71,000 documents, we only use a subset of 5,000 documents for our experiments (DE-Silver-5k).

We performed a human validation of the silver standard by manually revising 50 randomly selected documents by two reviewers. The reviewers reported a 14,2% of errors. While most errors come from the annotation itself, there are some errors that originate from the automatic insertion of synthetic names:

- Gender mismatch: *don Maria Carmen Rodriguez Soldevilla* is tagged as "given name - female", when in the text the subject is clearly a man, however the automatic insertion resulted in a female name.

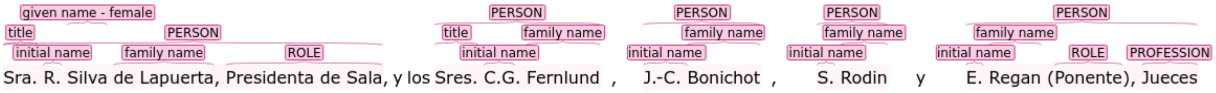
- Streets tagged as people: *la calle Maria Carmen Morales Garcia* is tagged as a PERSON entity, whereas it should be tagged as ADDRESS.

While we are aware of this issue, it is important to note that our final aim is to detect sensitive entities susceptible to be anonymised, therefore, even if the automatic preannotation was missing the right tag, it was targeting the correct tokens.

5. Experiments and Results

To validate the quality of the datasets, we fine-tuned several pre-trained language models on the different datasets and evaluated the obtained models. Thus, we performed a set of wide range of experiments by fine-tuning and evaluating on different combinations of fine-tuning and evaluation data and pre-trained models. In total, we performed 48 different experiments.

The entity detection model follows a sequence-labelling paradigm based on Transformers. The texts are split into tokens, which in turn are encoded into contextual-word-embeddings. Each of these embeddings pass through two classification heads, one for the level 1 entity types (the more coarse-grained, e.g. PERSON or ADDRESS) and other for the level 2 entity types (the more fine-grained, e.g. "Family name"



Sra. R. Silva de Lapuerta, Presidenta de Sala, y los Sres. C.G. Fernlund , J.-C. Bonichot , S. Rodin y E. Regan (Ponente), Jueces

Figure 1: Annotation interface in INCePTION.

	CPP		EUR-Lex		DE-Silver-5k/ DE-Gold	
Level 1 Ents	Train	Test	Train	Test	Train	Test
O	37,384	6,867	59,868	26,148	10,035,412	15,225
Address	118	16	117	45	43,600	92
Amount	52	12	10	28	21,400	40
Date	152	35	357	216	113,710	151
Person	456	117	365	171	48,298	84
Organisation	112	6	222	28	36,085	20
Time	27	5	0	0	1,485	1
Vehicle	4	0	5	5	1,758	0
	CPP		EUR-Lex		DE-Silver-5k/ DE-Gold	
Level 2 Ents	Train	Test	Train	Test	Train	Test
O	38,110	6,959	60,984	26,544	10,338,734	15,553
City	99	13	19	4	26,790	51
Country	2	0	100	39	8,100	2
Place	0	0	0	4	0	1
Postcode	2	0	0	0	574	0
Street	39	10	0	0	2,860	1
Territory	6	0	2	0	10,573	32
Unit	51	12	10	28	20,841	34
Value	53	12	10	28	21,452	42
Day	125	34	274	156	103,160	137
Month	135	33	285	159	103,476	135
Standard Abbreviation	16	1	0	0	983	1
Year	115	34	357	214	85,583	127
Licence Plate Number	2	0	0	0	309	0
Model	2	0	0	0	255	0
Type	0	0	5	5	784	0
Age	23	3	0	8	790	0
Ethnic Category	0	0	2	0	22	0
Family Name	335	60	169	121	36,605	56
Given Name - Female	101	17	6	5	20,127	26
Given Name - Male	192	74	17	8	16,651	37
ID Document Number	46	2	0	0	2,255	0
Initial Name	0	0	140	22	444	0
Marital Status	0	0	1	0	4	0
Nationality	11	1	0	0	353	2
Profession	0	0	11	0	15	0
Role	30	8	23	7	1,274	0
Title	162	24	64	18	7,054	5
Url	8	0	0	0	1	0

Table 3: Level 1 and Level 2 entity distribution across datasets. ’O’ refers to non-annotated tokens.

or ”City”). Figure 2 shows a high level diagram of the model.

The transformer model is trained along with the classification heads, i.e. its weights are updated during the training. Since the transformer model function is to encode the text into meaningful embeddings, in general, the better the pre-trained language-model, the better the results.

We evaluate our resources with different pre-trained language models, both multilingual and monolingual, and both general and domain-specific. The pre-trained models we have used are all available on the HuggingFace Hub[6]. They are the following:

[6] https://huggingface.co/

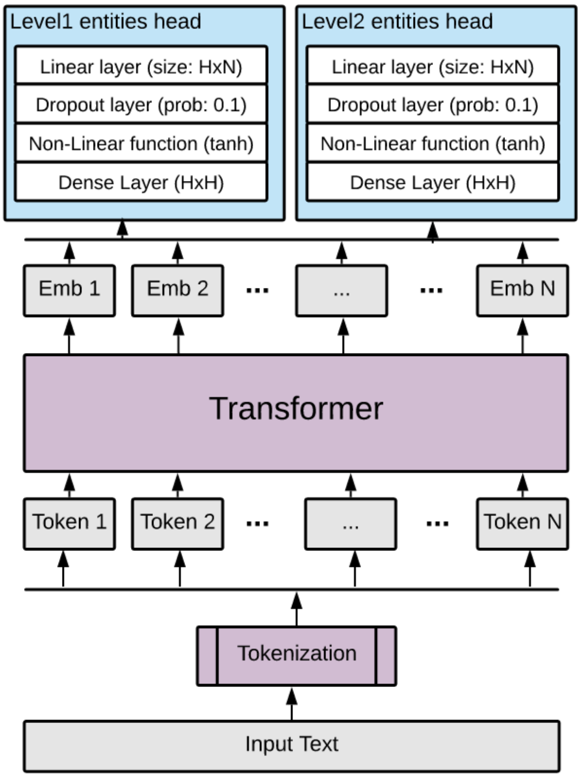


Figure 2: Model diagram. The text is encoded using a Transformer model, and each resulting token embedding (size H) is projected to the N classes of each entities level.

- **mBERT**: the multilingual version of BERT (Devlin et al., 2019). It was trained on monolingual data from the corresponding Wikipedia of the different languages.

  base-multilingual-cased

- **BETO**: a pre-trained BERT model solely trained on Spanish data proposed by Canete et al. (2020). The training data comes from Wikipedia and OPUS (Tiedemann, 2012).

  bert-base-spanish-wwm-cased

- **Spanish RoBERTa-b**: a pre-trained RoBERTa-base model trained on a total of 570GB sourced from the National Library of Spain, member of the MarIA Spanish language models' family (Gutiérrez-Fandiño et al., 2022).

  roberta-base-bne

- **RoBERTaLex**: a pre-trained RoBERTa-base model trained on legal domain corpora, including the resources described in this work, which represent 15,92% of the total training data (Gutiérrez-Fandiño et al., 2021).

  RoBERTaLex

We use F1 to report our results in Table 4.
As it is well known, NERC is a downstream task which usually obtains high results. This can be observed in the scores obtained, which range from 0.714 to 0.981. These results can be analysed in terms of the data used to fine-tune, the data used for evaluation and the underlying pre-trained models.
Regarding fine-tuning data, as could be expected, the best scores are obtained when the distribution of the fine-tuning and test set are the same (e. g. train on CPP and evaluate on CPP). Still, we can see how the highest scores overall are obtained by merging two datasets, namely CPP and EUR-Lex, which was possible because they have been annotated using the same schema. The combination of the datasets seems to contribute positively to the task, even if they belong to two different distributions.
In terms of evaluation datasets, the highest scores are obtained when evaluating with the CPP dataset. CPP is the shortest test set and the one that contains less entities, as shown in Table 2, which makes it the easiest test set. On the other hand, evaluating the DE-Gold test

	Evaluation Data											
	EUR-Lex				CPP				DE-Gold			
Fine-tuning Data	mB	B	R	RL	mB	B	R	RL	mB	B	R	RL
EUR-Lex	0.955	0.956	0.962	0.949	0.834	0.862	0.893	0.837	0.772	0.813	0.849	0.714
CPP	0.874	0.893	0.890	0.805	0.972	0.979	0.980	0.960	0.868	0.862	0.874	0.829
EUR-Lex + CPP	0.957	0.955	0.956	0.933	0.980	0.981	0.981	0.964	0.875	0.847	0.869	0.802
DE-Silver-5k	0.944	0.928	0.721	0.861	0.974	0.972	0.896	0.925	0.877	0.876	0.762	0.845

Fine-tuning Dataset	Avg. F1
EUR-Lex	0.866
CPP	0.899
EUR-Lex + CPP	**0.925**
DE-Silver-5k	0.882

Evaluation Dataset	Avg. F1
EUR-Lex	0.909
CPP	**0.937**
DE-Gold	0.833

Model	Avg. F1
mB	0.907
B	**0.910**
R	0.886
RL	0.869

Table 4: F1 Results for NERC on different evaluation datasets, with varying sizes of fine-tuning data and four different pre-trained language models: mBERT (mB), BETO (B), RoBERTa-b (R), RoBERTaLex (RL). We also report the average performance across all experiments, per fine-tuning dataset, evaluation dataset and model.

	mBERT	BETO	RoBERTa-b
mBERT			
BETO	0.8221		
RoBERTa-b	0.3516	0.2699	
RoBERTaLex	**0.0003**	**0.0004**	0.4308

Table 5: P Value for the models' performance across all experiments.

set produces the lowest scores across the board, probably because there is no manually annotated data of the same distribution.

Regarding to which pre-trained model works best, all models provide simlilar results. On average, we find that mBERT and BETO are the best performing systems in terms of higher scores across test sets. RoBERTaLex, intriguingly, is the worst performing, although the differences among scores is often minimal. mBERT, in particular, is able to get the most out of the automatically annotated DE-Silver corpus, compared to the other foundation models.

To further investigate the subtle differences in performance of the tested models, we conduct significance tests with the obtained results. We perform a two-tailed paired T-Test for every pair of systems. From the results in Table 5, we gather that RoBERTaLex is the only model that performs signfificantly worse than mBERT and BETO, since the P Value in these two cases is lower than 0.05.

6. A Real Use Case: The Spanish Ministry of Justice

Motivation During the course of the MAPA project, the Spanish Ministry of Justice (MoJ) manifested interest in testing our anonymisation tool. The MoJ has already in place a protocol to manually remove sensitive data from the documents that need to be shared among administrations or be published, so that privacy is preserved. In the manual process, the only automation they use are MS Word macros, therefore having access to a full automatic anonymisation tool was of great interest to them.

When developing anonymisation tools tailored to specific real-world use cases the problem of needing to deal with training data containing sensible information arises. The solution is similar to what we have described for the corpus Dictámenes del Estado, that is:

(i) identify personal information

(ii) annotate entities

(iii) replace original entities with syntetic entities, while preserving the coherence and structure of the original data.

Dataset Since, the MoJ was already using human annotators to anonymise their documents, they were able to prepare a manually annotated corpus where all the entities had already been manually replaced by fake entities. The corpus consisted of 120 documents annotated at sentence-level consisting of judicial resolutions, including court rulings, court orders and decrees. The source of the documents was balanced among 4 jurisdictions (civil, contentious-administrative, criminal and social) and 3 different Spanish provinces (Valladolid, Toledo and Murcia), in order to avoid geographical and thematic biases. The dataset statistics are shown in Table 6.

	Train	Test
Sents	1,222	216
Tokens	44,527	8,117
Tags	21,612	3,985
Entities	3,469	635

Table 6: Statistics of the MoJ dataset. 'Tags' refers to the individual tokens that are annotated, while 'Ents' refers to whole entities.

Annotation The MoJ dataset used a more restricted set of entities adapted to the needs of the use case, therefore the annotation scheme consisted only of one level hierarchy and contained the following tags: PEOPLE, ROLE, ORGANISATION, PLACE, DATE, LAW, CURRENCY, ID NUMBER, and CODE. The entity distribution is shown in table 7.

Entity	Train	Test
O	29,988	5,415
CODE	302	52
CURRENCY	86	27
DATE	209	40
ID NUMBER	47	9
LAW	186	32
ORGANISATION	426	80
PEOPLE	877	154
PLACE	187	38
ROLE	1,118	197

Table 7: Entity distribution of the MoJ dataset. 'O' refers to non-annotated tokens.

Model	MoJ
mBERT	0.937
BETO	0.934
RoBERTa-b	0.942
RoBERTaLex	0.934

Table 8: F1 results for NERC on the MoJ dataset using four different pre-trained language models

Results Results for the data provided by the Ministry of Justice are shown in table 8. For this experiment, we have trained and tested using the same distribution. All models have a similar performance with excellent results. RoBERTa-b reports the best scores, only by a difference of 0.05 points with mBERT and closely followed by a tie between BETO and RoBERTaLex.

7. Discussion

We have run a set of experiments with different pre-trained models to put to test the robustness and quality of datasets of the legal domain that we have gathered and annotated. Results of the experiments show that the performance of the models is good and consistent and validates the quality of the presented datasets. As could be expected, when the same distribution is used to fine-tune and evaluate, results are higher, confirming the convenience of model adaptation and customization. An interesting part of our evaluation exercise is to assess the suitability of the different foundation or pre-trained models: one multilingual model (mBERT), two general-purpose monolingual models (BETO and RoBERTa-b) and one in-domain trained model (RoBERTaLex). The results of our experiments confirm the benefits of multilingual pre-trained models, since we observe that mBERT performs very well across all text-specific tasks. We can thus conclude that, provided good-quality annotated data for fine-tuning, multilingual models are more than enough to obtain adequate results.

Regarding quality of the annotations versus amount of data, our results show that the combination of two manually annotated small datasets (EUR-Lex and CPP) results in better performance across all tests and outperforms the DE-Silver-5k experiments, with the expected exception of the DE-Gold test set.

Results on the DE-Gold test are proof of the usefulness of automatically annotated syntetic data as training corpus (DE-Silver), since the best scores are precisely obtained with this corpus. This demonstrates the potential and usefulness of this type of data, whenever there's no manually annotated data available.

Our experiments show that results benefit from fine-tuning the models with texts that are closer to the test (results of EUR-Lex and DE-Gold), but they also show that if the training corpus is too small (CPP) then adding an additional annotated corpus that is in-domain but not the same text type, such as EUR-Lex, can boost the scores.

Finally, domain adaptation is a tricky issue for any downstream task as oftentimes it is hard to define a domain. The fact that the domain-specific model RoBERTaLex gives significantly poorer results may be attributed to this. All the datasets presented here can be considered to belong to the general legal domain, but the type of content in each of them differs considerably.

8. Conclusions and Future Work

In this paper we describe four Spanish datasets in the legal domain, with annotations for entities that carry sensitive information. The aim of these datasets is to train and/or evaluate anonymisation tools capable of removing sensitive information from documents.

We have introduced the different datasets, explaining their differences and annotation process. The manual annotation on three of the datasets has been performed using the INCEpTION annotation tool. One dataset has been automatically annotated and a sample has been manually revised, with an error rate of 14,2%. To validate the quality of the datasets, we have split them into training and test sets, and built domain-adapted models, which we have then evaluated in a NERC task in a variety of scenarios. We have experimented with several pre-trained Transformer-based models, one multilingual, two general-domain monolingual and one in-domain monolingual, to assess the impact of the foundation model in the task at hand.

We have also tested the methodology with a real use case and proved that our approach can be applied successfully to other scenarios.

Regarding future work, it is worth further investigating the balance between good-quality manually annotated data with varying sizes of synthetic, automatically annotated data as a resource for fine-tuning pre-trained models for downstream tasks. We would also like to continue investigating the task of domain adaptation by exploring in more detail domain-specific models.

The resources presented in this paper are available in open repositories[7], except for the CPP dataset, which is available upon request.

Acknowlegements

This work was funded by the MAPA CEF project.

9. Bibliographical References

Bambroo, P. and Awasthi, A. (2021). Legaldb: Long distilbert for legal document classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4. IEEE.

Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559.*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[7] `https://elrc-share.eu/repository/browse/mapa-anonymization-package-spanish/b550e1a88a8311ec9c1a00155d026706687917f92f64482587c6382175dffd76/`

Francopoulo, G. and Schaub, L.-P. (2020). Anonymization for the gdpr in the context of citizen and customer relationship management and nlp. In *workshop on Legal and Ethical Issues (Legal2020)*, pages 9–14. ELRA.

Gianola, L., Ajausks, Ē., Arranz, V., Gibert, O. d., and Melero, M. (2020). Automatic removal of identifying information in official eu languages for public administrations: The mapa project. In *33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020): proceedings, Dec 2020, Brno, Prague, Czech Republic*, volume 334, pages 223–226. IOS Press.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Gonzalez-Agirre, A., and Villegas, M. (2021). Spanish legalese language model and corpora. *arXiv preprint arXiv:2110.12201*.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Armentano-Oller, C., Rodriguez-Penagos, C., Gonzalez-Agirre, A., and Villegas, M. (2022). Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Ji, S., Mittal, P., and Beyah, R. (2017). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys and Tutorials*, 19(2):1305–1326.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Lima-López, S., Perez, N., García-Sardiña, L., and Cuadros, M. (2020). Hitzalmed: Anonymisation of clinical text in spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7038–7043.

Mao, J. and Liu, W. (2019). Hadoken: a bert-crf model for medical document anonymization. In *IberLEF@ SEPLN*, pages 720–726.

Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodriguez, H., Martin, J. L., Villegas, M., and Krallinger, M. (2019). Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@SEPLN*.

Medlock, B. (2006). An introduction to nlp-based textual anonymisation. In *LREC*, pages 1051–1056. Citeseer.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

10. Language Resource References

Baisa, V., Michelfeit, J., Medveď, M., and Jakubíček, M. (2016). European union language resources in sketch engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803.

Samy, D., Arenas-García, J., and Pérez-Fernández, D. (2020). Legal-es: A set of large scale resources for spanish legal text processing. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 32–36.

Taranilla, R. (2012). La justicia narrante: Un estudio sobre el discurso de los hechos en el proceso penal. Aranzadi.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

A. Appendix

A.1. Annotation Scheme

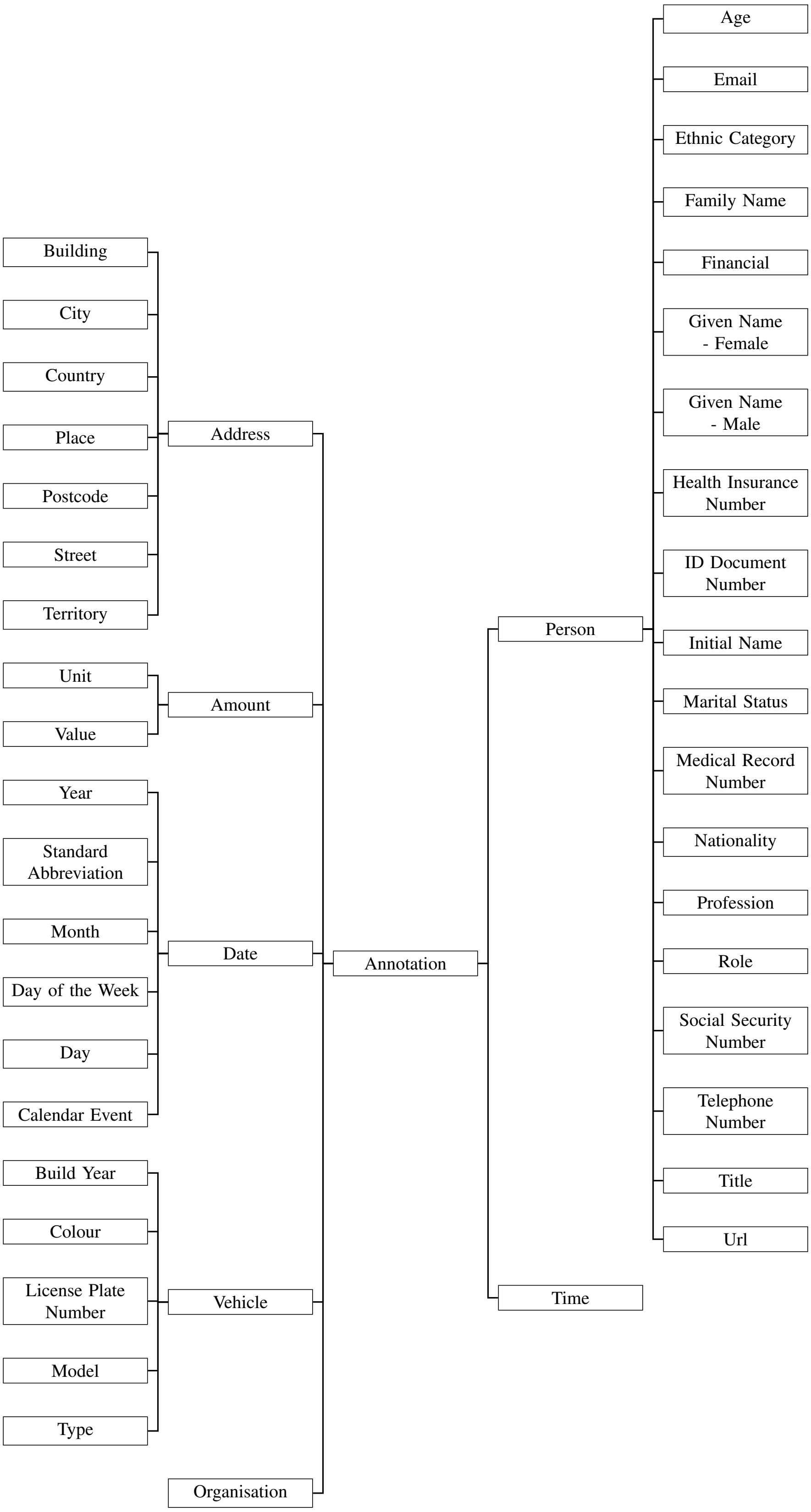


Figure 3: Annotation Scheme for the Mapa Project

B. Fine-tuning Hyperparameters

Hyper-parameter	Value
Optimizer	AdamW
Learning Rate	2E-5
Learning Rate Decay	Linear
Warmup	2 epochs
Batch Size	8
Max. Training Epochs	200
Early stopping patience	50

Table 9: Hyper-parameters used for fine-tuning the NERC models. All the trained models share the same training configuration.