Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4486–4494
Marseille, 11–16 May 2020

Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT

Aitor García-Pablos, Naiara Perez, Montse Cuadros
SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Donostia/San-Sebastián, 20009, Spain
{agarciap, nperez, mcuadros}@vicomtech.org

Abstract

Massive digital data processing provides a wide range of opportunities and benefits, but at the cost of endangering personal data privacy. Anonymisation consists in removing or replacing sensitive information from data, enabling its exploitation for different purposes while preserving the privacy of individuals. Over the years, a lot of automatic anonymisation systems have been proposed; however, depending on the type of data, the target language or the availability of training documents, the task remains challenging still. The emergence of novel deep-learning models during the last two years has brought large improvements to the state of the art in the field of Natural Language Processing. These advancements have been most noticeably led by BERT, a model proposed by Google in 2018, and the shared language models pre-trained on millions of documents. In this paper, we use a BERT-based sequence labelling model to conduct a series of anonymisation experiments on several clinical datasets in Spanish. We also compare BERT to other algorithms. The experiments show that a simple BERT-based model with general-domain pre-training obtains highly competitive results without any domain specific feature engineering.

Keywords: Anonymisation, De-identification, PHI, Clinical Data, BERT

1. Introduction

During the first two decades of the 21st century, the sharing and processing of vast amounts of data has become pervasive. This expansion of data sharing and processing capabilities is both a blessing and a curse. Data helps build better information systems for the digital era and enables further research for advanced data management that benefits the society in general. But the use of this very data containing sensitive information conflicts with private data protection, both from an ethical and a legal perspective.

There are several application domains on which this situation is particularly acute. This is the case of the medical domain (Abouelmehdi et al., 2018). There are plenty of potential applications for advanced medical data management that can only be researched and developed using real data; yet, the use of medical data is severely limited –when not entirely prohibited– due to data privacy protection policies. One way of circumventing this problem is to anonymise the data by removing, replacing or obfuscating the personal information mentioned, as exemplified in Table 1. This task can be done by hand, having people read and anonymise the documents one by one. Despite being a reliable and simple solution, this approach is tedious, expensive, time consuming and difficult to scale to the potentially thousands or millions of documents that need to be anonymised.

For this reason, numerous of systems and approaches have been developed during the last decades to attempt to automate the anonymisation of sensitive content, starting with the automatic detection and classification of sensitive information. Some of these systems rely on rules, patterns and dictionaries, while others use more advanced techniques related to machine learning and, more recently, deep learning. Given that this paper is concerned with text documents (e.g. medical records), the involved techniques are related to Natural Language Processing (NLP). When using NLP approaches, it is common to pose the problem of document anonymisation as a sequence labelling problem, i.e. classifying each token within a sequence as being sensitive information or not. Further, depending on the objective of the anonymisation task, it is also important to determine the type of sensitive information (names of individuals, addresses, age, sex, etc.).

The anonymisation systems based on NLP techniques perform reasonably well, but are far from perfect. Depending on the difficulty posed by each dataset or the amount of available data for training machine learning models, the performance achieved by these methods is not enough to fully rely on them in certain situations (Abouelmehdi et al., 2018). However, in the last two years, the NLP community has reached an important milestone thanks to the appearance of the so-called Transformers neural network architectures (Wolf et al., 2019). In this paper, we conduct several experiments in sensitive information detection and classification on Spanish clinical text using BERT (from 'Bidirectional Encoder Representations from Transformers') (Devlin et al., 2019) as the base for a sequence labelling approach. The experiments are carried out on two datasets: the MEDDOCAN: *Medical Document Anonymization* shared task dataset (Marimon et al., 2019), and NUBES (Lima et al., 2019), a corpus of real medical reports in Spanish. In these experiments, we compare the performance of BERT with other machine-learning-based systems, some of which use language-specific features. Our aim is to evaluate how good a BERT-based model performs without language nor domain specialisation apart from the training data labelled for the task at hand.

The rest of the paper is structured as follows: the next section describes related work about data anonymisation in general and clinical data anonymisation in particular; it also provides a more detailed explanation and background about the Transformers architecture and BERT. Section 3. describes the data involved in the experiments and the systems

original	Paciente de **64 años** operado de una hernia el **12/01/2016** por **la Dra Lopez**
example 1	Paciente de **XXXXXXX** operado de una hernia el **XXXXXXXXXX** por **XXXXXXXXXXX**
example 2	Paciente de **[-AGE-]** operado de una hernia el **[--DATE--]** por **[--DOCTOR--]**
example 3	Paciente de **59 años** operado de una hernia el **05/06/2019** por **el Dr Sancho**

Table 1: Anonymization examples of "64-year-old patient operated on a hernia on the 12/01/2016 by Dr Lopez"; sensitive data and their substitutions are highlighted in bold.

evaluated in this paper, including the BERT-based system; finally, it details the experimental design. Section 4. introduces the results for each set of experiments. Finally, Section 5. contains the conclusions and future lines of work.

2. Related Work

The state of the art in the field of Natural Language Processing (NLP) has reached an important milestone in the last couple of years thanks to deep-learning architectures, increasing in several points the performance of new models for almost any text processing task.

The major change started with the Transformers model proposed by Vaswani et al. (2017). It substituted the widely used recurrent and convolutional neural network architectures by another approach based solely on self-attention, obtaining an impressive performance gain. The original proposal was focused on an encoder-decoder architecture for machine translation, but soon the use of Transformers was made more general (Wolf et al., 2019). There are several other popular models that use Transformers, such as Open AI's GPT and GPT2 (Radford et al., 2019), RoBERTa (Liu et al., 2019) and the most recent XLNet (Yang et al., 2019); still, BERT (Devlin et al., 2019) is one of the most widespread Transformer-based models.

BERT trains its unsupervised language model using a Masked Language Model and Next Sentence Prediction. A common problem in NLP is the lack of enough training data. BERT can be pre-trained to learn general or specific language models using very large amounts of unlabelled text (e.g. web content, Wikipedia, etc.), and this knowledge can be transferred to a different downstream task in a process that receives the name *fine-tuning*.

Devlin et al. (2019) have used fine-tuning to achieve state-of-the-art results on a wide variety of challenging natural language tasks, such as text classification, Question Answering (QA) and Named Entity Recognition and Classification (NERC). BERT has also been used successfully by other community practitioners for a wide range of NLP-related tasks (Liu and Lapata, 2019; Nogueira and Cho, 2019, among others).

Regarding the task of data anonymisation in particular, anonymisation systems may follow different approaches and pursue different objectives (Cormode and Srivastava, 2009). The first objective of these systems is to detect and classify the sensitive information contained in the documents to be anonymised. In order to achieve that, they use rule-based approaches, Machine Learning (ML) approaches, or a combination of both.

Although most of these efforts are for English texts – see, among others, the i2b2 de-identification challenges (Uzuner et al., 2007; Stubbs et al., 2015), Dernoncourt et al. (2016), or Khin et al. (2018)–, other languages are also attracting growing interest. Some examples are Mamede et al. (2016) for Portuguese and Tveit et al. (2004) for Norwegian. With respect to the anonymisation of text written in Spanish, recent studies include Medina and Turmo (2018), Hassan et al. (2018) and García-Sardiña (2018). Most notably, in 2019 the first community challenge about anonymisation of medical documents in Spanish, MEDDOCAN[1] (Marimon et al., 2019), was held as part of the IberLEF initiative. The winners of the challenge –the Neither-Language-nor-Domain-Experts (NLNDE) (Lange et al., 2019)– achieved F1-scores as high as 0.975 in the task of sensitive information detection and categorisation by using recurrent neural networks with Conditional Random Field (CRF) output layers.

At the same challenge, Mao and Liu (2019) occupied the 8th position among 18 participants using BERT. According to the description of the system, the authors used BERT-Base Multilingual Cased and an output CRF layer. However, their system is $\sim$3 F1-score points below our implementation without the CRF layer.

3. Materials and Methods

The aim of this paper is to evaluate BERT's multilingual model and compare it to other established machine-learning algorithms in a specific task: sensitive data detection and classification in Spanish clinical free text. This section describes the data involved in the experiments and the systems evaluated. Finally, we introduce the experimental setup.

3.1. Data

Two datasets are exploited in this article. Both datasets consist of plain text containing clinical narrative written in Spanish, and their respective manual annotations of sensitive information in BRAT (Stenetorp et al., 2012) standoff format[2]. In order to feed the data to the different algorithms presented in Section 3.2., these datasets were transformed to comply with the commonly used BIO sequence representation scheme (Ramshaw and Marcus, 1999).

3.1.1. NUBES-PHI

NUBES (Lima et al., 2019) is a corpus of around 7,000 real medical reports written in Spanish and annotated with negation and uncertainty information. Before being published, sensitive information had to be manually annotated and replaced for the corpus to be safely shared. In this article,

[1] http://temu.bsc.es/meddocan/

[2] https://brat.nlplab.org/standoff.html

we work with the NUBES version prior to its anonymisation, that is, with the manual annotations of sensitive information. It follows that the version we work with is not publicly available and, due to contractual restrictions, we cannot reveal the provenance of the data. In order to avoid confusion between the two corpus versions, we henceforth refer to the version relevant in this paper as NUBES-PHI (from 'NUBES with Personal Health Information').
NUBES-PHI consists of 32,055 sentences annotated for 11 different sensitive information categories. Overall, it contains 7,818 annotations. The corpus has been randomly split into train (72%), development (8%) and test (20%) sets to conduct the experiments described in this paper. The size of each split and the distribution of the annotations can be consulted in Tables 2 and 3, respectively.

	train	dev	test
# sentences	23,079	2,565	6,411
# tokens	379,401	41,936	107,024
vocabulary	25,304	7,483	12,750
# annotations	5,562	677	1,579

Table 2: Size of the NUBES-PHI corpus

The majority of sensitive information in NUBES-PHI are temporal expressions ('Date' and 'Time'), followed by healthcare facility mentions ('Hospital'), and the age of the patient. Mentions of people are not that frequent, with physician names ('Doctor') occurring much more often than patient names ('Patient'). The least frequent sensitive information types, which account for ∼10% of the remaining annotations, consist of the patient's sex, job, and kinship, and locations other than healthcare facilities ('Location'). Finally, the tag 'Other' includes, for instance, mentions to institutions unrelated to healthcare and whether the patient is right- or left-handed. It occurs just 36 times.

	train		dev		test	
	#	%	#	%	#	%
Date	2,165	39	251	37	660	41
Hospital	1,012	18	105	16	275	17
Age	701	13	133	20	200	13
Time	608	11	63	9	155	10
Doctor	486	9	44	6	134	8
Sex	270	5	35	5	71	4
Kinship	158	3	20	3	44	3
Location	71	1	10	1	19	1
Patient	48	1	5	1	11	1
Job	31	1	3	0	9	1
Other	12	0	8	1	16	1
Total	5,562	100	677	100	1,579	100

Table 3: Label distribution in the NUBES-PHI corpus

3.1.2. The MEDDOCAN corpus

The organisers of the MEDDOCAN shared task (Marimon et al., 2019) curated a synthetic corpus of clinical cases enriched with sensitive information by health documentalists. In this regard, the MEDDOCAN evaluation scenario could be said to be somewhat far from the real use case the technology developed for the shared task is supposed to be applied in. However, at the moment it also provides the only public means for a rigorous comparison between systems for sensitive health information detection in Spanish texts.
The size of the MEDDOCAN corpus is shown in Table 4. Compared to NUBES-PHI (Table 2), this corpus contains more sensitive information annotations, both in absolute and relative terms.

	train	dev	test
# documents	500	250	250
# tokens	360,407	138,812	132,961
vocabulary	26,355	15,985	15,397
# annotations	11,333	5,801	5,661

Table 4: Size of the MEDDOCAN corpus

The sensitive annotation categories considered in MEDDOCAN differ in part from those in NUBES-PHI. Most notably, it contains finer-grained labels for location-related mentions –namely, 'Address', 'Territory', and 'Country'–, and other sensitive information categories that we did not encounter in NUBES-PHI (e.g., identifiers, phone numbers, e-mail addresses, etc.). In total, the MEDDOCAN corpus has 21 sensitive information categories. We refer the reader to the organisers' article (Marimon et al., 2019) for more detailed information about this corpus.

3.2. Systems

Apart from experimenting with a pre-trained BERT model, we have run experiments with other systems and baselines, to compare them and obtain a better perspective about BERT's performance in these datasets.

3.2.1. Baseline

As the simplest baseline, a sensitive data recogniser and classifier has been developed that consists of regular-expressions and dictionary look-ups. For each category to detect a specific method has been implemented. For instance, the Date, Age, Time and Doctor detectors are based on regular-expressions; Hospital, Sex, Kinship, Location, Patient and Job are looked up in dictionaries. The dictionaries are hand-crafted from the training data available, except for the Patient's case, for which the possible candidates considered are the 100 most common female and male names in Spain according to the Instituto Nacional de Estadística (INE; *Spanish Statistical Office*).

3.2.2. CRF

Conditional Random Fields (CRF) (Lafferty et al., 2001) have been extensively used for tasks of sequential nature. In this paper, we propose as one of the competitive baselines a CRF classifier trained with sklearn-crfsuite[3] for Python 3.5 and the following configuration: algorithm = `lbfgs`; maximum iterations = `100`; c1 = c2 = `0.1`; all transitions = `true`; optimise = `false`. The features extracted from each token are as follows:

[3]https://sklearn-crfsuite.readthedocs.io

– prefixes and suffixes of 2 and 3 characters;
– the length of the token in characters and the length of the sentence in tokens;
– whether the token is all-letters, a number, or a sequence of punctuation marks;
– whether the token contains the character '@';
– whether the token is the start or end of the sentence;
– the token's casing and the ratio of uppercase characters, digits, and punctuation marks to its length;
– and, the lemma, part-of-speech tag, and named-entity tag given by ixa-pipes[4] (Agerri et al., 2014) upon analysing the sentence the token belongs to.

Noticeably, none of the features used to train the CRF classifier is domain-dependent. However, the latter group of features is *language* dependent.

3.2.3. spaCy

spaCy[5] is a widely used NLP library that implements state-of-the-art text processing pipelines, including a sequence-labelling pipeline similar to the one described by Strubell et al. (2017). spaCy offers several pre-trained models in Spanish, which perform basic NLP tasks such as Named Entity Recognition (NER). In this paper, we have trained a new NER model to detect NUBES-PHI labels. For this purpose, the new model uses all the labels of the training corpus coded with its context at sentence level. The network optimisation parameters and dropout values are the ones recommended in the documentation for small datasets[6]. Finally, the model is trained using batches of size 64. No more features are included, so the classifier is language-dependent but not domain-dependent.

3.2.4. BERT

As introduced earlier, BERT has shown an outstanding performance in NERC-like tasks, improving the start-of-the-art results for almost every dataset and language. We take the same approach here, by using the model BERT-Base Multilingual Cased[7] with a Fully Connected (FC) layer on top to perform a fine-tuning of the whole model for an anonymisation task in Spanish clinical data. Our implementation is built on PyTorch[8] and the PyTorch-Transformers library[9] (Wolf et al., 2019). The training phase consists in the following steps (roughly depicted in Figure 1):

1. **Pre-processing:** since we are relying on a pre-trained BERT model, we must match the same configuration by using a specific tokenisation and vocabulary. BERT also needs that the inputs contains special tokens to signal the beginning and the end of each sequence.

2. **Fine-tuning:** the pre-processed sequence is fed into the model. BERT outputs the contextual embeddings that encode each of the inputted tokens. This embedding representation for each token is fed into the FC linear layer after a dropout layer (with a 0.1 dropout probability), which in turn outputs the logits for each possible class. The cross-entropy loss function is calculated comparing the logits and the gold labels, and the error is back-propagated to adjust the model parameters.

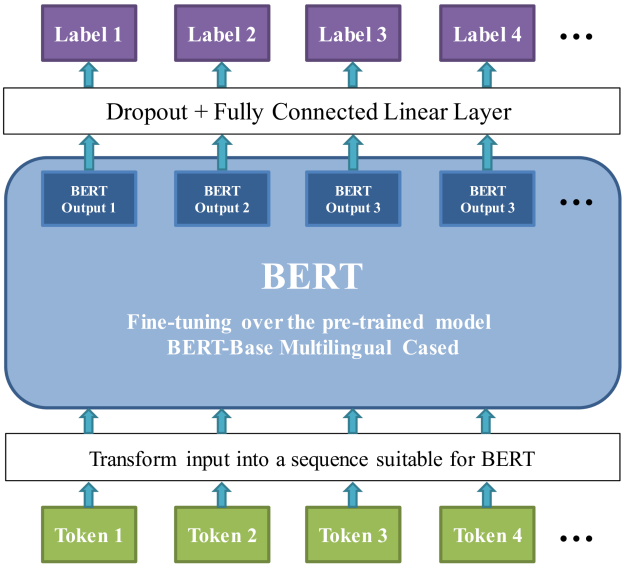


Figure 1: Pre-trained BERT with a Fully Connected layer on top to perform the fine-tuning

We have trained the model using an AdamW optimiser (Loshchilov and Hutter, 2019) with the learning rate set to 3e-5, as recommended by Devlin et al. (2019), and with a gradient clipping of 1.0. We also applied a learning-rate scheduler that warms up the learning rate from zero to its maximum value as the training progresses, which is also a common practice. For each experiment set proposed below, the training was run with an early-stopping patience of 15 epochs. Then, the model that performed best against the development set was used to produce the reported results. The experiments were run on a 64-core server with operating system Ubuntu 16.04, 250GB of RAM memory, and 4 *GeForce RTX 2080* GPUs with 11GB of memory. The maximum sequence length was set at 500 and the batch size at 12. In this setting, each epoch –a full pass through all the training data– required about 10 minutes to complete.

3.3. Experimental design

We have conducted experiments with BERT in the two datasets of Spanish clinical narrative presented in Section 3.1. The first experiment set uses NUBES-PHI, a corpus of real medical reports manually annotated with sensitive information. Because this corpus is not publicly available, and in order to compare the BERT-based model to other related published systems, the second set of experiments uses the MEDDOCAN 2019 shared task competition dataset. The following sections provide greater detail about the two experimental setups.

3.3.1. Experiment A: NUBES-PHI

In this experiment set, we evaluate all the systems presented in Section 3.2., namely, the rule-based baseline, the CRF

[4]https://ixa2.si.ehu.es/ixa-pipes
[5]https://spacy.io
[6]https://spacy.io/usage/training
[7]https://github.com/google-research/bert
[8]https://pytorch.org
[9]https://github.com/huggingface/transformers

	Detection			Classification (relaxed)			Classification (strict)		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
baseline	0.853	0.469	0.605	0.779	0.429	0.553	0.721	0.396	0.512
CRF	**0.968**	0.921	0.944	**0.952**	0.907	0.929	**0.941**	0.896	0.918
spaCy	0.964	0.938	0.951	0.942	0.852	0.895	0.942	0.852	0.895
BERT	0.952	**0.979**	**0.965**	0.938	**0.963**	**0.950**	0.925	**0.950**	**0.937**

Table 5: Results of Experiment A: NUBES-PHI

classifier, the spaCy entity tagger, and BERT. The evaluation comprises three scenarios of increasing difficulty:

Detection - Evaluates the performance of the systems at predicting whether each token is sensitive or non-sensitive; that is, the measurements only take into account whether a sensitive token has been recognised or not, regardless of the BIO label and the category assigned. This scenario shows how good a system would be at obfuscating sensitive data (e.g., by replacing sensitive tokens with asterisks).

Classification (relaxed) - We measure the performance of the systems at predicting the sensitive information type of each token –i.e., the 11 categories presented in Section 3.1.1. or 'out'. Detecting entity types correctly is important if a system is going to be used to replace sensitive data by fake data of the same type (e.g., random people names).

Classification (strict) - This is the strictest evaluation, as it takes into account both the BIO label and the category assigned to each individual token. Being able to discern between two contiguous sensitive entities of the same type is relevant not only because it is helpful when producing fake replacements, but because it also yields more accurate statistics of the sensitive information present in a given document collection.

The systems are evaluated in terms of micro-average precision, recall and F1-score in all the scenarios.

In addition to the scenarios proposed, a subject worth being studied is the need of labelled data. Manually labelled data is an scarce and expensive resource, which for some application domains or languages is difficult to come by. In order to obtain an estimation of the dependency of each system on the available amount of training data, we have retrained all the compared models using decreasing amounts of data –from 100% of the available training instances to just 1%. The same data subsets have been used to train all the systems. Due to the knowledge transferred from the pre-trained BERT model, the BERT-based model is expected to be more robust to data scarcity than those that start their training from scratch.

3.3.2. Experiment B: MEDDOCAN

In this experiment set, our BERT implementation is compared to several systems that participated in the MEDDOCAN challenge: a CRF classifier (Perez et al., 2019), a spaCy entity recogniser (Perez et al., 2019), and NLNDE (Lange et al., 2019), the winner of the shared task and current state of the art for sensitive information detection and classification in Spanish clinical text. Specifically, we include the results of a domain-independent NLNDE model (S2), and the results of a model enriched with domain-specific embeddings (S3). Finally, we include the results obtained by Mao and Liu (2019) with a CRF output layer on top of BERT embeddings. MEDDOCAN consists of two scenarios:

Detection - This evaluation measures how good a system is at detecting sensitive text spans, regardless of the category assigned to them.

Classification - In this scenario, systems are required to match exactly not only the boundaries of each sensitive span, but also the category assigned.

The systems are evaluated in terms of micro-averaged precision, recall and F-1 score. Note that, in contrast to the evaluation in Experiment A, MEDDOCAN measurements are entity-based instead of tokenwise. An exhaustive explanation of the MEDDOCAN evaluation procedure is available online[10], as well as the official evaluation script[11], which we used to obtain the reported results.

4. Results

This section describes the results obtained in the two sets of experiments: NUBES-PHI and MEDDOCAN.

4.1. Experiment A: NUBES-PHI

Table 5 shows the results of the conducted experiments in NUBES-PHI for all the compared systems. The included baseline serves to give a quick insight about how challenging the data is. With simple regular expressions and gazetteers a precision of 0.853 is obtained. On the other hand, the recall, which directly depends on the coverage provided by the rules and resources, drops to 0.469. Hence, this task is unlikely to be solved without the generalisation capabilities provided by machine-learning and deep-learning models.

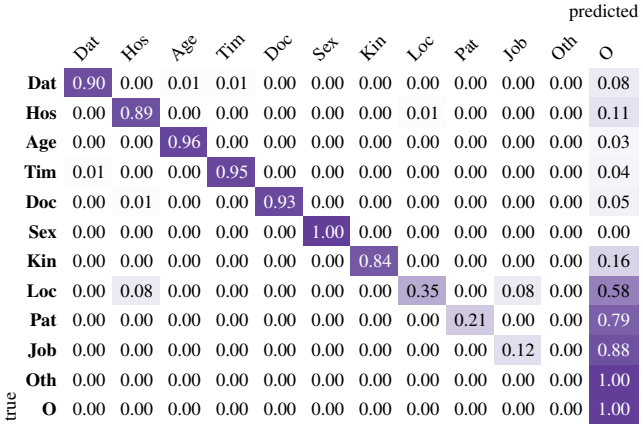
Regarding the detection scenario –that is, the scenario concerned with a binary classification to determine whether each individual token conveys sensitive information or not–, it can be observed that BERT outperforms its competitors. A fact worth highlighting is that, according to these results, BERT achieves a precision lower than the rest of the systems (i.e., it makes more false positive predictions); in exchange, it obtains a remarkably higher recall. Noticeably, it

[10] http://temu.bsc.es/meddocan/index.php/evaluation/

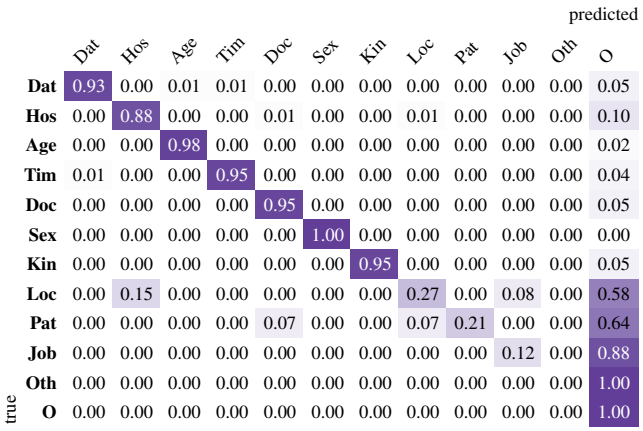
[11] https://github.com/PlanTL-SANIDAD/MEDDOCAN-Evaluation-Script

reaches a recall of 0.979, improving by more than 4 points the second-best system, spaCy.

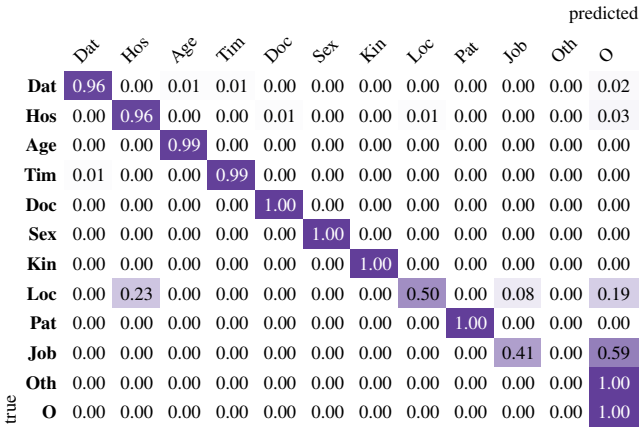
The table also shows the results for the relaxed metric that only takes into account the entity type detected, regardless of the BIO label (i.e., ignoring whether the token is at the beginning or in the middle of a sensitive sequence of tokens). The conclusions are very similar to those extracted previously, with BERT gaining 2.1 points of F1-score over the CRF based approach. The confusion matrices of the predictions made by CRF, spaCy, and BERT in this scenario



(a) CRF



(b) spaCy



(c) BERT

Table 6: Confusion matrices for the sensitive information classification task on the NUBES-PHI corpus

are shown in Table 6. As can bee seen, BERT has less difficulty in predicting correctly less frequent categories, such as 'Location', 'Job', and 'Patient'. One of the most common mistakes according to the confusion matrices is classifying hospital names as 'Location' instead of the more accurate 'Hospital'; this is hardly a harmful error, given that a hospital is actually a location. Last, the category 'Other' is completely leaked by all the compared systems, most likely due to its almost total lack of support in both training and evaluation datasets.

To finish with this experiment set, Table 5 also shows the strict classification precision, recall and F1-score for the compared systems. Despite the fact that, in general, the systems obtain high values, BERT outperforms them again. BERT's F1-score is 1.9 points higher than the next most competitive result in the comparison. More remarkably, the recall obtained by BERT is about 5 points above.

Upon manual inspection of the errors committed by the BERT-based model, we discovered that it has a slight tendency towards producing ill-formed BIO sequences (e.g, starting a sensitive span with 'Inside' instead of 'Begin'; see Table 7). We could expect that complementing the BERT-based model with a CRF layer on top would help enforce the emission of valid sequences, alleviating this kind of errors and further improving its results.

	Acudirá	a	la	Clínica	Marseille
true	O	O	B	**I**	I
predicted	O	O	B	**B**	I

(a) Example 1: "[The patient] will attend the Marseille Clinic"

	control	(	15	y	22	de	junio	)
true	O	O	B	O	**B**	I	I	O
predicted	O	O	B	O	**I**	I	I	O

(b) Example 2: "inspection (15 and 22 of june)"

	Niño	de	4	años	y	medio
true	**B**	O	B	I	**I**	**I**
predicted	**O**	O	B	I	**O**	**O**

(c) Example 3: "4 and a half years-old boy"

Table 7: BERT error examples (only BIO-tags are shown; differences between gold annotations and predictions are highlighted in bold)

Finally, Figure 2 shows the impact of decreasing the amount of training data in the detection scenario. It shows the difference in precision, recall, and F1-score with respect to that obtained using 100% of the training data. A general downward trend can be observed, as one would expect: less training data leads to less accurate predictions. However, the BERT-based model is the most robust to training-data reduction, showing an steadily low performance loss. With 1% of the dataset (230 training instances), the BERT-based model only suffers a striking 7-point F1-score loss, in contrast to the 32 and 39 points lost by the CRF and spaCy models, respectively. This steep performance drop stems to

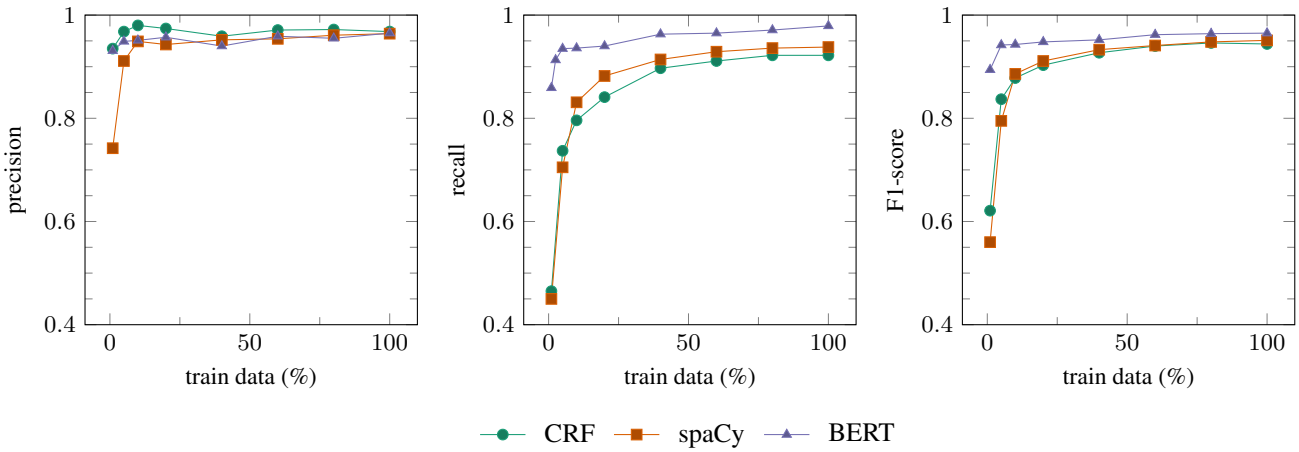


Figure 2: Performance decay with decreasing amounts of training data on the sensitive information detection task in the NUBES-PHI corpus

	Detection			Classification		
	Prec	Rec	F1	Prec	Rec	F1
CRF (Perez et al., 2019)	**0.977**	0.943	0.960	**0.971**	0.937	0.954
spaCy (Perez et al., 2019)	0.967	0.953	0.965	0.965	0.947	0.956
NLND S2 (Lange et al., 2019)	0.976	0.973	0.974	**0.971**	0.968	**0.970**
NLND S3 (Lange et al., 2019)	0.975	**0.975**	**0.975**	0.970	**0.969**	**0.970**
BERT + CRF (Mao and Liu, 2019)	0.968	0.919	0.943	0.965	0.912	0.937
BERT	0.973	0.972	0.972	0.968	0.967	0.967

Table 8: Results of Experiment B: MEDDOCAN

a larger extent from recall decline, which is not that marked in the case of BERT. Overall, these results indicate that the transfer-learning achieved through the BERT multilingual pre-trained model not only helps obtain better results, but also lowers the need of manually labelled data for this application domain.

4.2. Experiment B: MEDDOCAN

The results of the two MEDDOCAN scenarios –detection and classification– are shown in Table 8. These results follow the same pattern as in the previous experiments, with the CRF classifier being the most precise of all, and BERT outperforming both the CRF and spaCy classifiers thanks to its greater recall. We also show the results of Mao and Liu (2019) who, despite of having used a BERT-based system, achieve lower scores than our models. The reason why it should be so remain unclear.

With regard to the winner of the MEDDOCAN shared task, the BERT-based model has not improved the scores obtained by neither the domain-dependent (S3) nor the domain-independent (S2) NLNDE model. However, attending to the obtained results, BERT remains only 0.3 F1-score points behind, and would have achieved the second position among all the MEDDOCAN shared task competitors. Taking into account that only 3% of the gold labels remain incorrectly annotated, the task can be considered almost solved, and it is not clear if the differences among the systems are actually significant, or whether they stem from minor variations in initialisation or a long-tail of minor labelling inconsistencies.

5. Conclusions and Future Work

In this work we have briefly introduced the problems related to data privacy protection in clinical domain. We have also described some of the groundbreaking advances on the Natural Language Processing field due to the appearance of Transformers-based deep-learning architectures and transfer learning from very large general-domain multilingual corpora, focusing our attention in one of its most representative examples, Google's BERT model.

In order to assess the performance of BERT for Spanish clinical data anonymisation, we have conducted several experiments with a BERT-based sequence labelling approach using the pre-trained multilingual BERT model shared by Google as the starting point for the model training. We have compared this BERT-based sequence labelling against other methods and systems. One of the experiments uses the MEDDOCAN 2019 shared task dataset, while the other uses a novel Spanish clinical reports dataset called NUBES-PHI.

The results of the experiments show that, in NUBES-PHI, the BERT-based model outperforms the other systems without requiring any adaptation or domain-specific feature engineering, just by being trained on the provided labelled data. Interestingly, the BERT-based model obtains a remarkably higher recall than the other systems. High recall is a desirable outcome because, when anonymising sensible documents, the accidental leak of sensible data is likely to

be more dangerous than the unintended over-obfuscation of non-sensitive text.

Further, we have conducted an additional experiment on this dataset by progressively reducing the training data for all the compared systems. The BERT-based model shows the highest robustness to training-data scarcity, loosing only 7 points of F1-score when trained on 230 instances instead of 21,371. These observation are in line with the results obtained by the NLP community using BERT for other tasks. The experiments with the MEDDOCAN 2019 shared task dataset follow the same pattern. In this case, the BERT-based model falls 0.3 F1-score points behind the shared task winning system, but it would have achieved the second position in the competition with no further refinement.

Since we have used a pre-trained multilingual BERT model, the same approach is likely to work for other languages just by providing some labelled training data. Further, this is the simplest fine-tuning that can be performed based on BERT. More sophisticated fine-tuning layers could help improve the results. For example, it could be expected that a CRF layer helped enforce better BIO tagging sequence predictions. Precisely, Mao and Liu (2019) participated in the MEDDOCAN competition using a BERT+CRF architecture, but their reported scores are about 3 points lower than our implementation. From the description of their work, it is unclear what the source of this score difference could be. Further, at the time of writing this paper, new multilingual pre-trained models and Transformer architectures have become available. It would not come as a surprise that these new resources and systems –e.g., XLM-RoBERTa (Conneau et al., 2019) or BETO (Wu and Dredze, 2019), a BERT model fully pre-trained on Spanish texts– further advanced the state of the art in this task.

6. Acknowledgements

This work has been supported by Vicomtech and partially funded by the project DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER,UE).

7. Bibliographical References

Abouelmehdi, K., Beni-Hessane, A., and Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *Journal of Big Data*, 5(1):1–18.

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 3823–3828.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116*.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., and Szolovits, P. (2016). De-identification of Patient Notes with Recurrent Neural Networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

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.

García-Sardiña, L. (2018). Automating the anonymisation of textual corpora. Master's thesis, University of the Basque Country (UPV/EHU).

Hassan, F., Domingo-Ferrer, J., and Soria-Comas, J. (2018). Anonimización de datos no estructurados a través del reconocimiento de entidades nominadas. In *Actas de la XV Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2018)*.

Khin, K., Burckhardt, P., and Padman, R. (2018). A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation. In *28th Workshop on Information Technologies and Systems (WITS 2018)*.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289.

Lange, L., Adel, H., and Strötgen, J. (2019). NLNDE: The Neither-Language-Nor-Domain-Experts' Way of Spanish Medical Document De-Identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 671–678.

Lima, S., Perez, N., Cuadros, M., and Rigau, G. (2019). NUBes: a Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*.

Liu, Y. and Lapata, M. (2019). Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*.

Mamede, N., Baptista, J., and Dias, F. (2016). Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294.

Mao, J. and Liu, W. (2019). Hadoken: a BERT-CRF Model for Medical Document Anonymization. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 720–726.

Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodríguez, H., Lopez Martin, J. A., 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 *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 618–638.

Medina, S. and Turmo, J. (2018). Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nogueira, R. and Cho, K. (2019). Passage Re-ranking with BERT. *arXiv:1901.04085*.

Perez, N., García-Sardiña, L., Serras, M., and Del Pozo, A. (2019). Vicomtech at MEDDOCAN: Medical Document Anonymization. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, pages 696–703.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Ramshaw, L. A. and Marcus, M. P., (1999). *Natural Language Processing Using Very Large Corpora*, chapter 9, pages 157–176. Springer Netherlands.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pages 102–107.

Strubell, E., Verga, P., Belanger, D., and McCallum, A. (2017). Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2670–2680.

Stubbs, A., Kotfila, C., and Uzuner, O. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58(Supplement):S11 – S19.

Tveit, A., Edsberg, O., Røst, T., Faxvaag, A., Nytrø, Ø., Nordgård, T., Ranang, M., and Grimsmo, A. (2004). Anonymization of General Practioner Medical Records. In *Proceedings of the Second HelsIT Conference*.

Uzuner, O., Luo, Y., and Szolovits, P. (2007). Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–63.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Proceedings of the 31st International Conference on Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*.

Wu, S. and Dredze, M. (2019). Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems 32*, pages 5753–5763.