Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 33-42 recibido 18-10-2024 revisado 21-01-2025 aceptado 06-02-2025

Clinical Federated Learning for Private ICD-10 Classification of Electronic Health Records from Several Spanish Hospitals

Aprendizaje clínico Federado para la Clasificación en Base a CIE-10 de Historias Clínicas Electrónicas del Sistema Sanitario Español

Nuria Lebeña, Alberto Blanco, Arantza Casillas, Maite Oronoz, Alicia Pérez

HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)
nuria.lebena@ehu.eus

Abstract: A bottleneck in the Electronic Health Records (EHRs) classification according to the International Classification of Diseases (ICD) task is the challenge involved in getting large amounts of clinical Spanish documents for training efficient language models with private health data. The federated learning (FL) strategy enables the independent training of several models and the subsequent unification of each resulting model parameters to generate a unified model without the need to share sensitive data out of the clinical facilities. We analyse the feasibility of employing the federation strategy in Spanish in the context of an actual data division environment: data coming from two real hospitals from the Basque health system and generated in the same period. We also propose a method to further pre-train the language model (LM) in a federated manner. We apply our federated further pre-training method to the training of BETO and BERTmultilingual. Our findings clearly show that it is feasible to carry out federated learning for Spanish EHR classification using data spread across different hospitals. Moreover, the proposed LM further pre-training method steadily surpasses the results of the model without further pre-training.
Keywords: Clinical Natural Language Processing, Electronic Health Records in Spanish, International Classification of Diseases, Transformers Federated Learning.

Resumen: Una limitación en la clasificación de Registros Médicos Electrónicos (RMEs) según la Clasificación Internacional de Enfermedades (CIE) es el reto de conseguir grandes cantidades de documentos clínicos en castellano para entrenar modelos del lenguaje eficientes. El aprendizaje federado (FL) permite el entrenamiento independiente de varios modelos y la posterior unificación de los parámetros de cada modelo resultante para generar un modelo unificado sin necesidad de compartir datos sensibles fuera de las instalaciones clínicas. En este trabajo, analizamos la viabilidad de emplear la estrategia de federación en español en el contexto de una división de datos real: datos generados en el mismo periodo que provienen de dos hospitales reales del sistema de salud vasco. También proponemos un método para pre-entrenar el modelo de lenguaje (LM) de manera federada. Aplicamos este método de pre-entrenamiento federado al entrenamiento de BETO y BERTmultilingüe. Nuestros hallazgos muestran claramente que es factible llevar a cabo el aprendizaje federado para la clasificación de EHR en español utilizando datos distribuidos en diferentes hospitales. Además, la técnica propuesta de pre-entrenamiento federado mejora los resultados del modelo sin pre-entrenamiento adicional.
Palabras clave: Procesamiento del Lenguaje Natural, Registros Médicos Electrónicos en castellano, Clasificación Internacional de Enfermedades, Entrenamiento Federado de Transformers.

ISSN 1135-5948 DOI 10.26342/2025-74-2

1 Introduction

The issue of automatic Electronic Health Records (EHRs) classification according to The International Classification of Diseases (ICD) has received considerable critical attention in recent years (Xu et al., 2022). The International Classification of Diseases (ICD) encodes common clinical words and describes the range of diseases, disorders, injuries, and other related health conditions (Organization, 1993). This standard is used internationally with several purposes, among others, to aid information exchange, to generate mortality and morbidity statistics, to facilitate global information sharing, to monitor diseases, and also for billing purposes by insurance companies. Clinical coding is a mandatory task in numerous countries, including Spain, and is crucial in clinical documentation. Reading, understanding, and assigning ICD codes to each EHR is the task devoted to clinicians specifically-trained on clinical documentation and coding. Clinical documents, or EHRs, are documents that provide information about patients' health conditions; they contain information about past illnesses of the patient, lab test results, current health situation, patients response to treatment, illnesses of family members, etc. (van Aken et al., 2021). Expert coders are in charge of assigning the corresponding ICD codes to clinical documents, being in many cases doctors themselves who carry out this task. The code assignment is therefore an expensive and laborious task for healthcare systems. As a result, current research has focused on developing models capable of performing this task automatically (DeYoung et al., 2022; Yang et al., 2023). In this work, we deal with the automatic ICD assignation in Spanish EHRs (Barros et al., 2022).

Clinical document automatic classification according to the ICD is considered a Natural Language Processing (NLP) task. In recent years pre-trained language models (LM) have become the most widely used in these tasks (Grid, 2022). Performance in a range of NLP tasks, including ICD assignation, has significantly improved after pre-training language models on huge amounts of general domain unlabeled texts (Teng et al., 2023), even in Spanish (Gutiérrez-Fandiño et al., 2022). In recent years, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) has gained the most traction.

However, it is vital to train transformers with huge amounts of data. When working with documents from the medical domain getting clinical data is challenging due to difficulties with privacy and regulation. This is a major drawback in the ICD assignation task as there is a great lack of data to train the models. Moreover, when dealing with other languages rather than English, the issue worsens. In Spanish and other languages in which the available resources are scarce and hard to obtain, federation opens a door towards enhancing the extreme multi-label classification models by means of multi-lingual language models and sharing parameters across countries without sharing sensitive data. In this work we focus on Spanish clinical documents.

Federated learning (FL) enables the independent training of several models (i.e., one training procedure per data source or hospital) and the subsequent unification of each resulting model to generate a unified model without the need for sensitive information sharing. That is, this training strategy allows to train of independent models in different data sources without the need to centralize the data in a single device; only the updated weights and the parameters of the model are shared. This technique allows training models in a fragmented way, closer to how the data is distributed.

In the clinical domain, the data of the patients can be fragmented in different hospitals and in different data silos inside a hospital, for instance, distributed by service (traumatology, neurology...). Having data from a vast number of different sources favors the model training since it is possible to feed the model with concrete organizational vocabulary and language structures. This results in a model that is much more generalized than one that was trained merely using data from limited sources.

In this work we explore the federated learning strategy in the extreme multi-label classification of EHRs according to ICD using clinical documents in Spanish. More information on the system we wish to create and how its performance is measured is provided in Appendix I. We focus on a real data division, and we employ thousands of documents coming from two hospitals from the Basque health system (Osakidetza). We also analyse a federated further pre-training of BERT with a small subset of Spanish unannotated

clinical documents.

2 Related work

Previous studies have produced systems for federated classification applied to NLP tasks. In (Lin et al., 2022), they build a platform on FL that can assist with a variety of NLP tasks, including text generation, question answering, sequence tagging and language modelling. There has been only limited work on federated NLP in the clinical domain, none aimed at the extreme multi-label classification of ICDs. For instance, (Liu, Dligach, and Miller, 2019) carry out automatic phenotyping, a clinical NLP task that attempts to identify a group of patients that meet a predetermined set of standards. Others perform clinical Named Entity Recognition (Tian et al., 2022).

These studies do not carry out such a complex task as ours, and they outline that the trend is to divide the data into different silos artificially. To the best of our knowledge, we are the first to tackle this task in a federated manner, focusing on a real data division. The research that has been provided so far shows that the findings of federated learning presuppose a loss of respect for those of the centralized model.

Previous research shows that LMs can be further pre-trained by being fed millions of texts. For example, the Bio-BERT-Spanish model (Grid, 2022) is pre-trained with a clinical corpus of 5,157,902 free-text entries extracted from Chilean waiting list referrals. Moreover, the bsc-bio-es and bsc-bio-ehr-es models are trained on a biomedical-clinical corpus in Spanish, which was collected from various sources (Carrino et al., 2022). The use of these pre-trained models in the biomedical domain shows an improvement in classification results (De la Iglesia I. et al., 2023; Solarte-Pabón et al., 2023), also using a small set of data to further pre-train the LM models has shown improved results (Canete et al., 2020; Collado-Montañez, Martín-Valdivia, and Martínez-Cámara, 2025).

Recent studies have carried out the task of further pre-training in a federated manner (Tian et al., 2022; Liu and Miller, 2020). Nevertheless, in languages other than English, it is difficult to find alternative sources of massive data related to the medical domain. Therefore, we propose to carry out a federated LM further pre-training with a small subset of unannotated clinical notes coming from the same source as the annotated medical records. The aim is to check whether feeding the LM with a small amount of unannotated EHRs in a federated manner could improve the performance of federated classification and reduce the gap with centralized classification.

3 Methods

3.1 Materials

In this work, we employ a set of health records from Osakidetza, the Basque Health System. Indeed, there are two different hospitals involved, which will be referred to as *Hospital 1* and *Hospital 2*. We are focusing on a true case, with the set of EHRs naturally divided by hospital by contrast to the antecedents simulating the partition from a single source of data. Moreover, we account for the actual differences in the amount of clinical notes among hospitals, as the clinical records were collected during the same period in both medical centers.

The employed EHRs are written in Spanish and codified according to ICD-10 with multiple ICD codes per EHR leading to a multi-label task. Each ICD-10 code comprises three to seven alphanumeric characters and follow a hierarchy varying the specificity. The first three characters of the code refer to the main diagnosis referred to as the **Main** ICD granularity throughout this document. The remaining characters are known as the non-essential modifiers, and complete the primary code by including details, such as the disease's laterality and severity. This code level is known as fully-specified ICD (for the sake of brevity shall be referred to as **Full**). We find it of interest to assess the performance of the systems on both levels of specificity, Main and Full, as suggested by previous works (Blanco, Pérez, and Casillas, 2021; Duarte et al., 2018).

In terms of content, the EHRs or clinical documents are unstructured documents that report the incidents that affect patients during their current admission. They consist of multiple sections, including but not limited to: chief complaint, history of present illness, past medical history, medications on admission, an overview of the patient's hospital course and discharge summary. They might also include lab results and results of tests performed during admission.

Table 1 shows the quantitative description of the Osa dataset. All documents involved in this study were originally anonymous, thus, the pre-processing undergone include simple text-cleaning, such as removal of stopwords and special characters. Both hospitals' datasets (H1 and H2) are close in terms of the amount of notes and vocabulary. From the table note that the EHRs were coded, on average, with more than 5 ICD codes out of 1,773 in the fully-specified view, this is, indeed, a multi-label classification task.

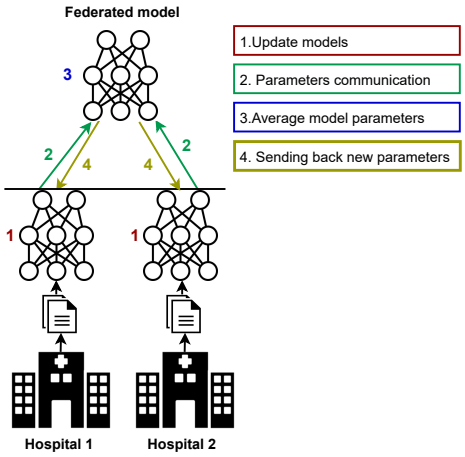
Concerning the labelling (ICD codes), we have obtained a subset of labels that are present in both hospitals, training parallel models with the same label-set. When taking into account the Main-class code granularity, the number of labels decreases by 58 % compared with the Fully-specified granularity. In terms of the average number of EHRs that have an ICD (avg. EHR in Table 1), there is an increase of 150 % when using the Main-class label-set.

In the further pre-training phase, an **unannotated dataset** is employed. The unannotated dataset is a collection of clinical documents extracted from the same source as the supervised datasets: a hospital. In this case, the aim of using this dataset is to determine whether the categorization process would be improved by the use of non-annotated documents unique to each facility.

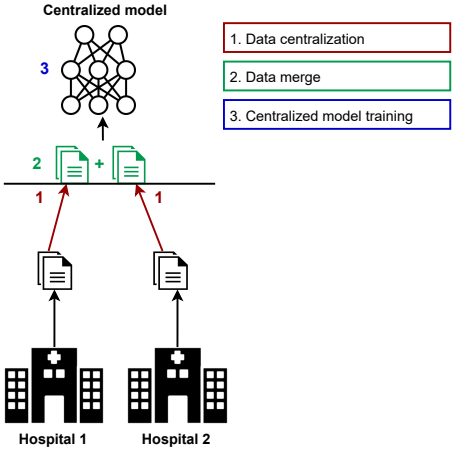
As the source is similar, the vocabulary of this dataset is close to the *Hospital 1* and *Hospital 2* datasets. 194,162 EHRs make up this unannotated dataset, which results in a vocabulary (number of distinct words) of 1,057,787 elements. To fit inside the BERT model's maximum input sequence length of 512 tokens, all the documents are shortened.

3.2 Federated Learning

There are several approaches to federated learning strategies. In our case, due to the structure of the data, we employ a model-central and horizontal federated learning approach. In a model-central federation, the distributed data is stored in each data source (the two hospitals in our work). When datasets have the same feature space but distinct samples, as in our case, horizontal federated learning is employed. Then, each model's parameters are employed to improve a central model via the use of an aggregation method (Rodríguez-Barroso et al., 2023). In Figure 1a we present a schema of our architecture in contrast to centralized architecture 1b



(a) Federated learning architecture



(b) Centralized learning architecture

Figura 1: Federated learning architecture and centralized learning architecture. In centralized learning, data is shared out the hospitals to train the model in a single data lake. Whereas in federated learning, each facility trains its own model while sharing the learned parameters to build a centralized model.

In a federated learning system, data silos train a model M_{fed} collectively as opposed to centralized data training, where the data is centralized in a data lake and the single model is trained using that data. Let $F = \{F_1, F_2, ..., F_k\}$ be a set of K data silos (hospitals in our case), each of them with a private dataset D_k of size n_k, where $k \in \{1, 2, ..., K\}$. Each of the silos F_k trains an independent model M_k with parameters Q_k. The goal is to train a collaborative global model M_{fed}, without divulging each facility's data D_k, as in (1).

	Documents		
	Notes	**Vocab**	**Avg W**
H1	13,507	89,422	767.61
H2	13,426	94,612	803.96
H1+H2	26,969	131,983	781.46

	Labels (ICDs)			
	Full: 1773		**Main: 743**	
	EHR/ ICD	**ICD/ EHR**	**EHR/ ICD**	**ICD/ EHR**
H1	38	5.11	97	5.32
H2	42	5.67	106	5.88
H1+H2	81	5.38	203	5.60

Tabla 1: Quantitative description of the **Osa** dataset hospitals (H1 and H2). The left table contains a quantitative description of the input. On the table on the right, the description of the output, the ICDs are described either as fully-specified diagnostic terms or as the main-class without non-essential modifiers leading to different cardinality (1,773 and 743 respectively). EHR/ICD indicates the number of clinical records in which an ICD is present on average; ICD/EHR represents the number of diagnoses covered by a clinical record on average.

$$L_k(Q_k) = \frac{1}{n_k} \sum_{(x_i,y_i)\in D_k} l(M_k(x_i;Q_k), y_i) \quad (1)$$

Where $l(\cdot,\cdot)$ is the Binary Cross-Entropy loss function, x_i the input features, and y_i the corresponding labels.

The central model M_{fed} is built by aggregating each individual model's (M_k) parameters. The process of weight update is known as communication and is carried out after a certain number of epochs (two in our case) in each individual model. Once the central weights are updated, they are returned to the independent models in order to keep training them in each device.

In communication, the weights are updated according to FedAvg mechanism: we average each individual model's (M_k) parameters by sample size, following expression (2). Where M^t_{fed} represents the centralized model at communication t and k is the number of data silos. n_k is the number of samples at the k^{th} silo, N is the total number of samples adding up all the silos, and Q^t_k represents the parameters learned from the k^{th} data silo (Konečnỳ et al., 2016). Note that expression (2) merely describes the process by which the weights of the network were updated during one communication of the training process; in a fully federated training process, this update occurs as many as time-iterations (t) involved.

$$M^t_{fed} = \sum_{k=1}^{2} \frac{n_k}{N} Q^t_k \quad (2)$$

In an attempt to promote scientific reproducibility, we made the source code available [1]

3.3 Federated LM further pre-training

In this work, we employ the Bidirectional Encoder Representations from Transformers (BERT (Devlin et al., 2018)) in its Spanish (BETO (Canete et al., 2020)), and multilingual (BERT multilingual (Pires, Schlinger, and Garrette, 2019)) variants. BERT is suitable for this task as it was designed to infer LMs as bidirectional representations from unannotated text, conveying information from both the left and right contexts. BETO is a BERT variant trained on a big Spanish corpus for Spanish NLP tasks. BERT multilingual is a BERT variant pre-trained on 104 languages, aimed at being useful in NLP tasks in languages other than English. We use and compare the performance of these two BERT variants as our documents are in Spanish. We are interested in determining whether the advantages offered by the federated LM further pre-training method are maintained even with diverse pre-training strategies.

Federated LM further pre-training entails improving the LM with a small set of unannotated corpora. This is done prior to training the classification head and is carried out by further pre-training BERT with unlabeled documents so that the language model is fitted to the specific data in the task.

Due to the scarcity of substantial clinical datasets appropriate to our task in Spanish, we employ a set of unannotated EHRs that come from the same hospitals, that is, being

[1] Web: `https://ixa2.si.ehu.eus/nlebena/federatedlearning/FederatedLearning.zip` Username: `federated` Password: `learning`

the source of the data the same as in the training data of the downstream task in each silo.

The federated further pre-training is carried out following expression (2), sharing BERT's parameters for each hospital with the centralized model.

We initially started to train parallel models (models with the same initial parameters) in both silos, that is, one with each hospital's dataset. We trained the model for 30 epochs, as suggested by previous research (Blanco, Pérez, and Casillas, 2021). We wanted to simulate continuous training with live data (as it is received), and we found out that the independent models made a more substantial improvement after two epochs rather than just one. We communicated the models once per two epochs with a total of 15 communications.

3.4 Federated head training classification

To face the extreme multi-label classification task, we fine-tuned BETO and BERT multilingual with a classification head. The classification head is a neural network-based module to which the contextual information previously retrieved by the transformer is passed. This classification module consists of a linear layer followed by a dropout layer and a Sigmoid activation function. In the resulting vector, each position is related to a label and indicates the estimated probability of that label being present in the given EHR. The architecture of the implemented module is presented in Figure 2

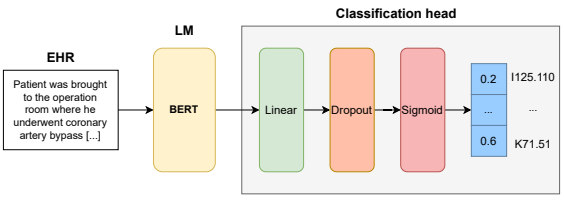


Figura 2: Classification model architecture.

We apply federated training to the classification head as in the federated LM further pre-training: with the same initial parameters in both silos, we begin to train parallel models. We then centralize and communicate with the models once every two epochs for a total of 15 communications. The update of the central model was carried out following expression (2). By contrast to the LM further pre-training phase, only the classification head weights are exchanged with the centralized model in the federated training stage.

4 *Experimental results*

The first research question was to assess if the predictive performance attained by each **individual** hospital could benefit from federated learning. To this end, we conducted a preliminary experiment without LM further pre-training with BERT, shown in Table 2 and focused on the label-set just with the Main granularity level.

Training strategy	P	R	F-1
Hospital 1	19.55	25.10	20.56
Hospital 2	21.65	26.70	23.19
Federated	26.31	32.81	27.85
Centralized	28.92	33.78	30.25

Tabla 2: Multi-label classification performance of the baselines (Hospital 1 and Hospital 2) are compared to their counterpart with joint effect training either as federated or centralized.

The findings demonstrate that employing a federated training strategy offers major improvements over training two independent models for each hospital. Comparing the federated and centralized results, the performance is close, being the federated approach weaker than the centralized and following the trend observed by previous works in other tasks (Kim et al., 2017; Luboshnikov and Makarov, 2021).

Next, we focused on determining if the **LM further pre-training** would be beneficial in the multi-label classification task. Additionally, the aim was to measure the performance loss between a federated and centralized classification. To this end, we have evaluated and compared four training methods for each model, combining the federation of the classification head and the LM further pre-training. We refer to centralized classification when the model is trained with the data from the two silos centralized in one device. We did not restrict ourselves to BERT and also assessed BETO, each at two granularity levels (fully-specified and main). The results of our experiments are shown in Table 3. The performance attained is explored varying label granularity: Table 3a shows the results attained for the fully-specified diagnostic term or full label granularity, while Ta-

ble 3b shows the results attained for the main label granularity.

Regarding the experiments without the LM- further pre-training, the average loss percentage between the centralized and federated performance is around 4.0 % using both BERT (4.35 % loss on average) and BETO (3.87 % loss on average).

Conducting a prior **LM further pre-training** results in an improvement of the federated model reducing the gap between federated and centralized approaches when federated further pre-training is not applied. In table 3, LM- further pre-training is therefore set to "Yes" if the LM model was enriched using the unannotated data. Both tables show how the LM further pre-training has an impact on the two BERT models: BERTmultilingual and BETO. The experiments were carried out with both levels of label granularity: full (Table 3a) and main (Table 3b), to see the performance in tasks of varying complexity. We can see that the enriched LMs steadily surpass the results of the models with further pre-training.

As disclosed in Table 1, the number of EHRs containing a given ICD, on average, is small; just 38 EHRs (out of 13,507) are devoted to each ICD on average. We were concerned about the fact that learning from a few samples is challenging, and so reveal the results reported in Table 3. Some authors restricted the datasets where some guarantee of repeatability is present (Berndorfer and Henriksson, 2017; Dermouche et al., 2016). Following the criteria of previous works, we have conducted parallel experiments. We merely focused on ICDs that were present in at least 1 % of the documents (we referred to this subset as Osa-1r). This setup led to the results presented in Table 4.

As in Table 3 the experiments with the Osa 1-r reduced set (Table 4) were conducted with full (Table 4(a)) and main (Table 4(b)) label granularity. The enriched LMs slightly exceed the results of the models with further pre-training, even with more code representativity (this trend was also observed in the experiments involving all the codes).

Comparing Tables 3 and 4, we noted that, as the number of clinical notes available per ICD increased, the ability of the models to learn increased, as well, in absolute values. We found that the ability of the federated LM further pre-training approach follows the same trend as the centralized training strategy, being beneficial in both scenarios. In terms of predictive ability, BETO's performance is better when predicting labels for the Osa dataset (Table 4), while BERT outperforms BETO when the label-set is reduced in Osa-1r dataset (Table 3).

5 Discussion

The current study found that the difference between federated and centralized training widens in tasks with more label repeatability. The gap between centralized and federated learning is about 12-13 %. Still, in domains where it is not possible to obtain data due to privacy constraints, federated training shows to be a good alternative. As in previous work, (Lin et al., 2022) our loss interval between a federated and centralized training, ranges between 6-14 %, proving that federated learning for clinical documents ICD classification in Spanish can be carried out.

As it is natural, we also found that overall results improve when the label-set and label granularity is reduced, getting the best performance when the difficulty is lower. Furthermore, concerning the differences between BERTmultilingual and BETO, we can see that the advantages of LM further pre-training are consistent in both LMs.

Although the performance of the model improves when reducing the label granularity, the gap between centralized and federated learning increased when reducing the number of labels. By reducing the universe of labels to only those most present there are more documents per label, being the prevalence of some of these labels unbalanced in each hospital. We can see that this is detrimental to the federated model concerning the centralized one. We hypothesize that if some labels are better learned in one hospital and not as well in the other one it generates some noise in the federated approach with respect to the centralized one. In an environment in which the amount of labels is bigger, this is not so much appreciated, as most labels tend to be underrepresented

Even so, the results of this study are consistent with the current state of the art for the classification of EHRs in Spanish. Specifically, our results are comparable by previous research having studied hierarchical classification and further pre-training with centralized training (Blanco, Pérez, and Casillas,

		BERT			BETO		
Training	**LM further pre-training**	**P**	**R**	**F1**	**P**	**R**	**F1**
F	No	12.61	18.94	14.01	12.75	17.02	14.07
	Yes	**13.80**	**20.72**	**15.33**	**13.96**	**18.64**	**15.40**
C	No	18.77	23.08	19.84	20.10	20.93	19.46
	Yes	**20.68**	**25.43**	**21.86**	**21.91**	**22.81**	**21.21**

(a) Performance predicting fully-specified ICDs

		BERT			BETO		
Training	**LM further pre-training**	**P**	**R**	**F1**	**P**	**R**	**F1**
F	No	26.31	32.81	27.85	27.34	33.87	28.83
	Yes	**26.81**	**33.42**	**28.37**	**27.87**	**34.52**	**29.38**
C	No	28.92	33.78	30.25	36.10	35.31	34.73
	Yes	**29.91**	**34.95**	**31.29**	**36.92**	**36.12**	**35.52**

(b) Performance predicting the ICD classes restricted to main granularity

Tabla 3: Multi-label classification performance of federated learning strategy (F) against the centralized learning strategy (C) with and without LM further pre-training. The results are given in terms of weighted average Precision (P), Recall (R) and F1-Score (F1) distinguishing two levels of granularity: a) Fully-specified ICD; b) Main ICD without non-essential modifiers.

		BERT			BETO		
Training	**LM further pre-training**	**P**	**R**	**F1**	**P**	**R**	**F1**
F	No	33.57	36.33	33.13	33.57	35.72	32.29
	Yes	**37.69**	**40.48**	**36.78**	**31.18**	**39.56**	**35.76**
C	No	47.91	46.87	46.96	46.70	40.65	43.09
	Yes	**52.49**	**51.35**	**51.44**	**50.17**	**46.68**	**48.99**

(a) Performance predicting fully-specified ICDs

		BERT			BETO		
Training	**LM further pre-training**	**P**	**R**	**F1**	**P**	**R**	**F1**
F	No	44.07	45.29	42.92	43.85	43.79	41.48
	Yes	**46.09**	**46.65**	**44.08**	**44.97**	**44.91**	**42.54**
C	No	56.88	55.27	55.70	53.65	48.14	50.37
	Yes	**57.75**	**56.11**	**56.54**	**57.97**	**55.58**	**56.30**

(b) Performance predicting the ICD classes restricted to main granularity

Tabla 4: Multi-label classification performance on the **Osa-1r**. Results of federated learning strategy (F) against the centralized learning strategy (C) with and without LM further pre-training are given. The performance is given at two levels of granularity of the ICD: a) Fully-specified ICD; b) Main ICD without non-essential modifiers. The results are given in terms of weighted average Precision (P), Recall (R) and F1-Score (F1).

2021). Moreover, our performance is also in line with other studies that have classified medical records in Spanish and have similar amounts of labels and training examples (Almagro et al., 2020; Mou and Ren, 2020).

5.1 Conclusion

The proposed federated LM further pre-training technique proves to improve the federated and centralized classification, making it appropriate to apply when data sharing is not possible. Using Osa unsupervised health records to centrally enrich an LM provides us with a 10 % performance improvement. The federated fine-tuning also show a improved performance in a 1-5 % range in both BETO and BERT.

Federating aids in enhancing the effectiveness of each silo in contrast to training them independently. Although centralized systems are superior, federated systems allow for improvement without sharing data (exchanging

models) which would allow getting more data to conduct the training process and eventually overcome the gap with the centralized classification. Nevertheless, our study has some limitations, our findings suggest that further research is needed before deploying federated training models into production. Future work should focus on existing models that were pre-trained with Spanish clinical data (Carrino et al., 2022; Grid, 2022). We also plan to assess the effect of federated further pre-training based on the size of the dataset available for pre-training.

Moreover, future work should assess the effects of federated training involving more hospitals and also segmenting the silos by clinical service or medical specialty. More broadly, research is also needed to determine whether having more data in each silo could improve a model trained with fewer centralized data.

Acknowledgment

This work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31); by Antidote PCI2020-120717-2 and LOTU TED2021-130398B-C22 funded by the MCIN/AEI /10.13039/501100011033 and by the European Union NextGenerationEU/ PRTR; by the Basque Government (IXA IT-1570-22, Predoctoral Grant PRE-2022-1-0069); and by EXTEPA within Misiones Euskampus 2.0.

References

Almagro, M., R. M. Unanue, V. Fresno, and S. Montalvo. 2020. ICD-10 Coding of Spanish Electronic Discharge Summaries: An Extreme Classification Problem. *IEEE Access*.

Barros, J., M. Rojas, J. Dunstan, and A. Abeliuk. 2022. Divide and conquer: An extreme multi-label classification approach for coding diseases and procedures in Spanish. In *Workshop on Health Text Mining and Information Analysis*, pages 138–147, December.

Berndorfer, S. and A. Henriksson. 2017. Automated diagnosis coding with combined text representations. *Stud Health Technol Inform*, 235:201–5.

Blanco, A., A. Pérez, and A. Casillas. 2021. Exploiting ICD Hierarchy for Classification of EHRs in Spanish Through Multi-Task Transformers. *IEEE JBHI*, 26.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR*, 2020:1–10.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Biomedical Language Processing*, pages 193–199. Association for Computational Linguistics, May.

Collado-Montañez, J., M.-T. Martín-Valdivia, and E. Martínez-Cámara. 2025. Data augmentation based on large language models for radiological report classification. *Knowledge-Based Systems*, 308:112745.

De la Iglesia I., M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, and A. Atutxa. 2023. An open source corpus and automatic tool for section identification in Spanish health records. *Journal of Biomedical Informatics*, 145:104461.

Dermouche, M., J. Velcin, R. Flicoteaux, S. Chevret, and N. Taright. 2016. Supervised topic models for diagnosis code assignment to discharge summaries. In *Intelligent Text Processing and Computational Linguistics*, pages 485–497. Springer.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

DeYoung, J., H.-C. Shing, L. Kong, C. Winestock, and C. Shivade. 2022. Entity anchored ICD coding. *arXiv preprint arXiv:2208.07444*.

Duarte, F., B. Martins, C. S. Pinto, and M. J. Silva. 2018. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *Journal of biomedical informatics*, 80:64–77.

Grid, E. L. 2022. Bio-bert-spanish. Version 1.0.0 (automatically assigned). [Model].

Source: European Language Grid. https://live.european-language-grid.eu/catalogue/ld/14256.

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

Kim, Y., J. Sun, H. Yu, and X. Jiang. 2017. Federated tensor factorization for computational phenotyping. In *Knowledge discovery and data mining*, pages 887–895.

Konečnỳ, J., H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *CoRR*, abs/1610.05492.

Lin, B. Y., C. He, Z. Ze, H. Wang, Y. Hua, C. Dupuy, R. Gupta, M. Soltanolkotabi, X. Ren, and S. Avestimehr. 2022. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In *NAACL 2022*, Seattle, United States, July. Association for Computational Linguistics.

Liu, D., D. Dligach, and T. Miller. 2019. Two-stage federated phenotyping and patient representation learning. In *Association for ComputationalLinguistics.*, volume 2019, page 283. NIH Public Access.

Liu, D. and T. Miller. 2020. Federated pretraining and fine tuning of BERT using clinical notes from multiple silos. *CoRR*, abs/2002.08562.

Luboshnikov, E. and I. Makarov. 2021. Federated learning in named entity recognition. *Recent Trends in Analysis of Images, Social Networks and Texts*, 1357:90.

Mou, C. and J. Ren. 2020. Automated ICD-10 code assignment of nonstandard diagnoses via a two-stage framework. *Artificial Intelligence in Medicine*, 108:101939.

Organization, W. H. 1993. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, volume 2. World Health Organization.

Pires, T., E. Schlinger, and D. Garrette. 2019. How multilingual is multilingual BERT? In *Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Rodríguez-Barroso, N., D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara. 2023. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173.

Solarte-Pabón, O., O. Montenegro, A. García-Barragán, M. Torrente, M. Provencio, E. Menasalvas, and V. Robles. 2023. Transformers for extracting breast cancer information from Spanish clinical narratives. *Artificial Intelligence in Medicine*, 143:102625.

Teng, F., Y. Liu, T. Li, Y. Zhang, S. Li, and Y. Zhao. 2023. A review on deep neural networks for icd coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375.

Tian, Y., Y. Wan, L. Lyu, D. Yao, H. Jin, and L. Sun. 2022. Fedbert: When federated learning meets pre-training. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–26.

van Aken, B., J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. A. Gers, and A. Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *EACL*, pages 881–893. Association for Computational Linguistics.

Xu, J., X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui. 2022. A survey of deep learning for electronic health records. *Applied Sciences*, 12(22):11709.

Yang, Z., S. Kwon, Z. Yao, and H. Yu. 2023. Multi-label few-shot icd coding as autoregressive generation with prompt. In *Conference on Artificial Intelligence*, volume 37.