Universidad del País Vasco (UPV/EHU)
Departamento de Electricidad y Electrónica

PhD dissertation

Extreme multi-label deep neural classification of Spanish health records according to the International Classification of Diseases

Alberto Blanco Garcés

2022

Universidad del País Vasco (UPV/EHU)
Departamento de Electricidad y Electrónica

Extreme multi-label deep neural classification of Spanish health records according to the International Classification of Diseases

This is the report of the thesis "Extreme multi-label deep neural classification of Spanish health records according to the International Classification of Diseases", written by Alberto Blanco Garcés under the supervision of Dr. Alicia Pérez and Dr. Arantza Casillas.

Getxo (2022).

Acknowledgments

This thesis was partially funded by the Spanish Ministry of Science and Innovation through the PROSAMED (TIN2016-77820-C3-1-R) and DOTT-HEALTH projects (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), the European Commission (FEDER), and the Basque Government (IXA IT-1343-19, Predoctoral Grant PRE-2019-1-0158). We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research. We thank Julián Salvador and Idoia Anso (Osakidetza) for providing the necessary support in the clinical domain.

Abstract

This work deals with clinical text mining, a field of natural language processing applied to the biomedical domain. The aim of this work is to automatise the medical coding task. Electronic health records (EHR) are documents that contain clinical information about the health of a patient. The medical diagnoses embodied in the EHRs are coded with respect to the International Classification of Diseases (ICD). Indeed, the ICD is the foundation for identifying international health statistics as well as the standard for reporting diseases and health conditions. From a machine learning perspective, the goal of this work is to solve an extreme multi-label text classification problem; each health record is assigned multiple ICD codes from a set of over 70,000 diagnostic terms. A significant amount of resources are devoted to medical coding, a laborious task that is currently done manually. The EHRs are extensive narratives and medical coders review the records written by physicians and assign the corresponding ICD codes. The texts are technical because the clinicians employ specialised medical jargon. However, the EHRs are also rich in abbreviations, acronyms, and spelling mistakes because clinicians document the records while engaged in actual clinical practice. To address the automatic classification of health records, we researched and developed a set of deep learning text classification techniques. Furthermore, clinical decisions are critical matters. Therefore, we investigated the interpretability of models to generate knowledgeable, accountable, and consistent predictions.

Contents

Document outline

This dissertation is organised as a compendium of publications on clinical multi-label text classification, the research focus of this thesis. The manuscript complies with the "thesis by published papers" regulation (also referred to as "thesis by compilation") from the University of the Basque Country (UPV/EHU) that dictates the format and the structure of the manuscript. This document must consist of at least three articles published in scientific journals featured in the latest list published by the Journal Citation Reports (JCR), Scopus, or the databases listed by the National Assessment Committee for Research Activities (CNEAI). At least one of these publications must belong to the first or second quartile of their category.

The overall structure of the document takes the form of two chapters. In the first part of Chapter 1, we introduce and motivate the proposed research. Then, the main objectives and research questions (RQs) are outlined in Section 1.1. After that, in Section 1.2, we present an overview of the works developed within the framework of the research. Then, in Sections 1.2.1-1.2.5, we delve into the 5 works that constitute the compilation. To end the chapter, Section 1.3 contextualises the results by summarising the best results and providing an overview of the progression of the developed approaches. The purpose of Chapter 2 is to reflect on the lessons learned, conclusions, and open research questions this thesis highlights. The full articles from the compilation can be found in Appendix A.

1 Introduction

Electronic health records (EHR) are the digital versions of medical records, containing a comprehensive compilation of facts pertinent to an individual's health history, covering aspects of the patient's past or present physical medical conditions, mental medical conditions, illnesses, and treatments. The records emphasise the specific events affecting the patient during the current episode of care. In short, EHRs are digital documents used to record patient medical data that health professionals require to provide adequate care.

The available volume of healthcare data has exploded with the adoption of EHRs; its systematic collection is critical to public health (Safran et al., 2007). The **International Classification of Diseases** (ICD) coding system (World Health Organization et al., 1975) maintained by the World Health Organization is the standard for diagnostic health information used for EHR coding. It is used worldwide for epidemiology, health management, and documentation purposes. Medical classification consists of transforming descriptions of medical diagnoses or procedures from EHRs into codes from medical standards. The ICD is designed as a medical classification system. Therefore, it provides the standard reference with tens of thousands of codes for conditions, signs, symptoms, abnormal findings, complaints, social circumstances, or external causes of diseases and injuries.

A growing body of literature documents the value of EHRs for advantageous secondary uses, such as analytics or predictive systems (Yadav et al., 2018). Nevertheless, to enable a broad set of secondary uses at a large scale, coded EHRs are necessary, i.e., it is required to extract structured data from the unstructured health records (Goenaga et al., 2021). Currently, in many

countries' health systems, EHRs are interpreted by experts and manually assigned diagnostic codes. The focus of this work is applying natural language processing (NLP) and machine learning (ML) methods to the clinical domain to automatically assign EHR codes. The **manual coding** of EHRs is cumbersome. Medical coders, expert clinicians trained on coding guidelines (Mujtaba et al., 2017), must search for critical information inside the lengthy, unstructured narratives texts of EHRs. Next, the expert coders must choose the codes to assign from a set of tens of thousands of labels, considering the anatomic location, etiology, severity, and laterality of the medical condition.

The health records are hundreds or thousands of words in length and include around a dozen ICDs that have to be chosen from a vast set of codes. Additionally, some of these ICDS are not expressly mentioned in the health record itself. Therefore, medical coding is challenging, time-consuming, error-prone, and expensive. Moreover, coders must keep abreast of changes as the ICD evolves. For example, the transition from ICD-9 to ICD-10 had a significant impact on coders and the medical coding system because of the differences between these versions. As a result, there is a lack of codified EHRs in health systems, even from developed countries (Sankoh and Byass, 2014); this system results in unnecessary expenditure and labor.

Natural language processing is gaining traction in clinical documentation services as a means of coping with the massive amounts of data transmitted by EHRs (Gu et al., 2020). The medical coding can be tackled as a **multi-label classification problem** in which the input to classify is the EHR and the output is a set of diagnostic codes from the ICD. The medical coding process would be remarkably enhanced with machine-learning-powered solutions, speeding it up and easing the adaptation to new ICD versions. ML would also diminish costs, liberate human capital, and improve coding consistency. Ideally, the process would be fully automated, but it could also facilitate the task as a Decision Support System, as there is evidence that NLP can assist expert coders (Zikos and DeLellis, 2018).

Figure 1.1 shows the structure of an ICD-10 code with the three levels of granularity applied throughout the works that compose this thesis. The "chapter" level keeps only the first character (i.e., the ICD chapter), allowing a maximum of 24 labels. The "block" level keeps the first three characters, leading to a maximum of 1.910 unique combinations. At the "full-code" level, the ICD comprises 71.704 labels. Note that the levels of granularity are just another way to look at the hierarchy of the ICD. Following the example from Figure 1.1, if solely the block part is kept ("M1A") and non-essential

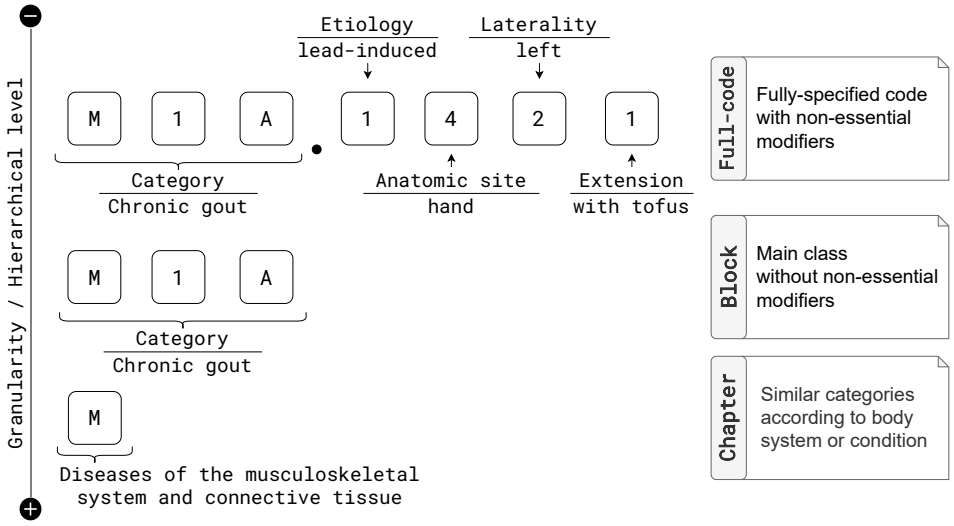


Figure 1.1: ICD-10 code structure unrolled at three levels of granularity. The diagnostic term "Lead-induced chronic gout in the left hand with tofus" is encoded as the "M1A.1421" full-code with all the non-essential modifiers.

modifiers are removed, the specificity is reduced to "chronic gout." Finally, when preserving only the first character, we get the ICD chapter, which groups diagnostic terms into similar categories according to body system or health condition. In this case, chapter "M" is indicated, which corresponds to "diseases of the musculoskeletal system and connective tissue."

EHRs are kept, along with their associated ICD codes, in digitalised healthcare systems. However, only a tiny amount of all EHRs are coded due to the arduous work that the task demands. Additionally, the lack of codified EHRs is relevant because ML, especially deep learning (DL) techniques, requires vast supervised corpora to train models (Sun et al., 2017). Figure 1.2 shows a sample EHR written in English, similar to those applied in our experiments. Precisely, we have mainly used the MIMIC dataset for English EHRs and the Osakidetza dataset for Spanish EHRs. We used intensive care unit discharge summaries from the Beth Israel Deaconess Medical Center and emergency services discharge summaries from Basque Country Health System. Table 1.1 shows a brief quantitative description of the EHRs from both datasets. Regarding ICD versions, we used ICD-9 codes when working with the MIMIC and ICD-10 codes with the Osakidetza dataset.

Chief complaint: fall, abdominal pain

History of Present Illness:

70 y/o man with h/o EtOH abuse[1], was found down intoxicated by EMS. Complained of having flu[2] for several days. One day prior to presentation, he had generalized abdominal pain[3]. No fever or chills. No prior episodes. He denied any blood, melena, dysuria.

Past Medical History:

- COPD[4]
- Hypertension[5]
- Hypercholesterolemia[6]
- Atrial Fibrillation[7] - anticoagulated on coumadin
- s/p coronary stent placement

Medications on Admission:
Ppropranolol 20 mg PO QID
Captorpil - ?unclear if he is on this, no dosage listed
Coumadin 6mg po qday

CT abd/pelvis: fatty enlarged liver[8], normal appendix, no free air

Brief Hospital Course:

The patient was admitted to the blue surgery service for treatment of pancreatits[9]. He was kept NPO and maintained on IVF. Empiric antibiotics were initiated. A Foley catheter and NG tube were placed. His stool tested positive for C. difficile[10] and Flagyl was prescribed.

He was subsequently transferred to the floor for further management where he continued to make good progress. He abdominal pain was minimal and generally relieve with Tylenol. He was up and ambulating with the use of a cane. Clinical symptoms resolved.

LABORATORY:

WBC-3.5*[11] RBC-5.99 HGB-12.4 HCT-39.4 MCV-99 MCH-31.2 MCHC-31.6 RDW-13.2
GLUCOSE-94 UREA N-33 CREAT-0.9 SODIUM-144POTASSIUM-3.7 CHLORIDE-105 TOTAL CO2-28

DISCHARGE SUMMARY:

- AGallstone pancreatitis[9]
- COPD[4], HCL[6], HTN[5], A Fib[7]

Discharge medications:

1. Acetaminophen 325 mg Tablet Sig: Two (2) Tablet PO Q6H as needed for pain, fever.
2. Docusate Sodium 100 mg Capsule Sig: One (1) Capsule PO BID: Hold for any loose stools.
3. Metronidazole 500 mg Tablet Sig: One (1) Tablet PO 3TID for 4 days. Disp:*12 Tablet(s)*

At the time of discharge was clinically stable and appropriate for discharge. He will stay off coumadin for three weeks because of the surgery, but then at that time he will restart anticoagulation. [**lastname **] will follow up in PC with Dr. [**Name**].

Gold standard ICD codes: F10.10[1], J10.1[2], R10.84[3], J44.9[4], I10[5], E78.00[6], I48.2[7], K76.0[8], K85.10[9], A04.7[10], D72.819[11]

Figure 1.2: Electronic health record sample. The pieces of text associated with each ICD code are indicated through the superscripts.

Following the EHR sample from Figure 1.2, we shall depict the main challenges derived from the nature of medical texts. First, the ICD codes are motivated by different types of mentions: i) An explicit mention of the standard term, meaning that the piece of text from the EHR matches the ICD standard description (i.e., "`generalized abdominal pain`"), ii) An explicit non-standard mention of the term, meaning that the piece of text from the EHR does not fully match the ICD standard description (i.e., "`flu`", "`hypertension`"), iii) An implicit mention, namely, when the diagnostic term is not expressly pointed out (i.e., a laboratory result such as the "`WBC-3.5*`", meaning that the white blood cells are below the normal range, leading to Leukocytopenia). Additionally, codes can be motivated by abbreviations or acronyms (i.e., "`HCL`", "`A Fib`"), or even by mentions with spelling mistakes (i.e., "`pancreatits`" instead of "pancreatitis"). Finally, negations must also be considered, since, if a diagnostic term is preceded by a negative particle, the corresponding ICD should not be coded (i.e., "`no fever`", "`denied blood`"). Furthermore, some labels are related, meaning that the presence or absence of a given label potentially influences any other label (i.e., "F10.19 – Alcohol abuse" promotes the presence of "K76.0 – Fatty liver"). Lastly, note that the EHRs occasionally contain differentiated sections (i.e., "Past Medical History" or "Discharge Summary"), but the present sections and header texts may differ notably across records because relevant sections differ between services and they are not stated equally by all physicians. In other words, it is unstructured text.

		Datasets	
		MIMIC	**Osakidetza**
$\mathcal{X}$	EHRs	55,172	26,969
	Vocab	137,207	379,121
	Words/EHR	1,399 ± 721	864 ± 415
$\mathcal{Y}$	Unique ICDs	6,918	5,541
	Avg. Card.	11.5	5.8

Table 1.1: Quantitative description of the EHRs from the main employed datasets, MIMIC and Osakidetza.

As shown in Figure 1.2, it is common to find spelling mistakes, abbreviations, or acronyms in EHRs because the reports are documented by physicians while engaged in practical work. For that reason, in conjunction

with the naturally complex and specialised jargon, the vocabulary (i.e., the number of unique words) escalates considerably. On the most extensive Osakidetza dataset, a set of 26,969 EHRs led to a vocabulary of over 379,121 unique words. Regarding the coded ICDs, the total number of codes is usually high, depending on the dataset size. From the 72,184 ICD-10 diagnostic terms, our largest Osakidetza label set contains 5,541 ICD-10 codes. There are 6,918 ICD-9 codes in the MIMIC dataset. Moreover, as observed in the sample EHR, the average number of ICD codes per document is also high—around 6 labels on average for the Osakidetza dataset, and over 11 in MIMIC. The vast number of labels and the natural distribution of medical conditions (i.e., a few highly prevalent conditions and a large set of rare conditions) produces class imbalance. For example, there are $3,552$ labels appearing in less than 1% of the 26,969 Osakidetza EHRs, while $1,885$ labels appear only once.

Multi-label classification problems are considered extreme when the cardinality of the label set is vast, such as the tens of thousands of diagnostic terms present in the ICD. The main challenge lies in the exponential label space that it implies. **Extreme multi-label text classification** (XMTC) serves as the framework for EHR classification. Moreover, when the labels are dependent on each other (e.g., there are inter-relationships among diseases), dependency-ignoring methods fail to predict label combinations coherently. However, research has shown that artificial intelligence (AI) systems in healthcare can significantly reduce the labour burden on health workers and even outperform human professionals at some tasks (McKinney et al., 2020). Therefore, applying machine learning to medical tasks such as EHR classification is an opportunity to increase healthcare efficiency and help reduce healthcare budget, which is constantly increasing in the United States as well as European countries (Dupor and Guerrero, 2021). Specifically, EHR classification according to the ICD opens the opportunity to apply data mining techniques to clinical data, which would advance fundamental tasks such as pharmaco-surveillance and the collection of morbidity and mortality statistics. It is also relevant for legal purposes, like billing for insurance companies and hospitals.

Internationally, the biomedical NLP research community has garnered significant interest. Horizon 2020 is the most significant research and innovation framework initiative from the European Union, with almost 80 billion Euros in available funding (European Commission, 2020). At the start of 2019, a total of 4,865 out of 20,877 projects were related to biomedical and

health research, with 94 countries contributing to at least one biomedical project (Gallo et al., 2021). In its National Strategy for Artificial Intelligence whitepaper, the Spanish government stated that the healthcare system is a strategic sector for promoting the research, development, and implementation of AI systems (Vicepresidencia Tercera, Gobierno de España, 2020). According to the study, the strategic health sector is in the top 4 sectors that will experience the most medium- and long-term impact from the implementation of AI systems. In this sense, the application of AI in health research will promote strategic projects that can lead to reform in and increase the efficiency of health care.

Biomedical research on NLP and ICD coding has piqued the interest of several international congresses and workshops. Since 2012, the CLEF eHealth Evaluation Lab and Workshop Series have been held annually, including several ICD coding challenges. In 2016, large-scale classification tasks were introduced (Névéol et al., 2016). Then, in the 2017 and 2018 editions, previous information extraction tasks were extended and the coding of death certificates with the ICD-10 was introduced (Névéol et al., 2017, 2018). The 2017 edition featured French and English languages, while Hungarian and Italian were introduced in 2018 to increase the focus on languages other than English. Non-technical summaries of animal experimentation are a kind of clinical document also annotated with ICD-10, which was the foundation of the 2019 CLEF eHealth Challenge (Dörendahl et al., 2019). Note that the length of death certificates and summaries of animal experimentation are hundreds of words shorter than discharge summaries, lessening the difficulty compared to our task. The 2020 edition (Goeuriot et al., 2020) focused on Spanish ICD-10 term coding for textual data collected from clinical records. Additionally, national and international programs such as the "Plan de Tecnologías del Lenguage" by the Spanish Ministry of Health (Gobierno de España, 2020) and the European Horizon 2020 (European Commission, 2020) program are driving the creation and publication of shared clinical datasets.

This thesis has been developed within the IXA group in the framework of the PROSA-MED (Díaz de Ilarraza Sánchez et al., 2017) and DOTT-HEALTH (Araujo et al., 2021) coordinated research projects, which aim to develop text-based technology to support diagnosis and disease prevention. The projects propose the automatic analysis of electronic health records to ascertain patterns, prevent errors, improve quality, reduce costs, and save time for the health services.

To summarise, several **challenges** arise naturally, some related to the

multi-label classification itself and others common to the clinical text mining in general. First, the genre of the text is relevant because of its implications in the availability of resources, vocabulary size scaling, writing style, and type of errors. In our case, we are working with electronic health records, which implies a scarcity of resources, a vast vocabulary, a narrative style, and an abundance of abbreviations as well as many orthographic and typographic errors. Furthermore, we focus mainly on EHRs written in Spanish—a language with scarce NLP and biomedical data resources compared to English. Moreover, the ICD classification is an instance of extreme multi-label classification (XMT), dealing with thousands of labels and thus highly inter-related label sets.

1.1 Objectives

This thesis aims to tackle the automatic coding of diagnostic terms present in free narrative, unstructured medical records according to the ICD. Therefore, we aim to conceive and develop machine learning models that aid in the automatic multi-label classification of EHRs according to the International Classification of Diseases, processing a narrative EHR to determine the appropriate diagnostic terms (i.e., categorising the record with the corresponding ICD codes). The following are some aspects we must accomplish: i) find an appropriate approximation to solve the task, a suitable family of methods (e.g., classical methods, deep learning, etc.), an adequate classifier, and a proper representation of the input—namely, the feature set, ii) build methods that generalise adequately and are not built ad-hoc to specific languages (i.e., multilingual models) or types of health records, iii) assess the influence of the input characteristics (i.e., the raw text from EHRs) and output (i.e., the set of labels) on the performance of the classifier and conceive ways to enhance the classification, adjusting the input and output, iv) explore the influence of the relationships among labels in the multi-label setting and develop techniques for predicting a consistent set of labels. Given this context, the main objective of this work and its corresponding sub-objectives are as follows:

Main objective: *Develop a method to automatically determine the diagnostic terms enclosed in an EHR according to the ICD*

- **Objective 1** *Develop versatile text classifiers for diagnostic term classification*: There are numerous kinds of health records, all of which have their own intrinsic characteristics that affect the output of the classifiers. We aim to develop classifiers adaptable enough to handle diverse health records such as discharge summaries, nursing notes, critical care unit admissions, or diagnostic impressions. Furthermore, it is valuable to develop classification models that can work with EHRs in various languages or from different medical specialties and hospitals (i.e., not tied to a specific language or service).

- **Objective 2** *Perform extreme multi-label text classification on lengthy documents*: The total number of labels on the ICD is exceptionally high. Even when the label set is bounded, we frequently work with high cardinality label sets of hundreds or thousands of labels. Additionally, our task involves classifying lengthy documents such as a complete medical histories or intensive care unit encounters. Therefore, we focus on increasing the capacity of the models to handle larger label sets and develop robust models, i.e., models that, when facing alterations in the input text, do not result in altering the predicted labels (Qayyum et al., 2020). As for the robustness of EHR classification, we interpret to not altering the predicted ICDs, even if the relevant information to detect the codes is scattered in records of hundreds or thousands of words (from 800 to 1,400 words on average).

- **Objective 3** *Predict explainable and coherent label sets of ICDs*: Modelling the label dependencies in document classification tasks is vital. The reason is that labels are tightly coupled due to the concurrent or mutually exclusive relationship. For example, some diseases tend to co-occur, such as diabetes and hypertension, while it is incoherent to simultaneously classify an EHR with an adult- and child-specific disease. Moreover, the ICD is arranged hierarchically, which is another matter to consider in the label relationships. We aim to develop interpretable classifiers that promote the prediction of related diagnostic terms while preventing the co-appearance of incompatible medical conditions.

Our **hypothesis** is, therefore, that NLP and deep learning classification techniques can mimic the process that expert human coders accomplish. To ascertain its validity, we must solve a natural language understanding (NLU) problem because expert coders examine, understand, and analyse the EHRs

to determine which explicit and implicit diagnostic terms are enclosed and assign the corresponding ICD codes. To illustrate the intricacies of solving an NLU task, consider the following example of an EHR fragment in natural language:

```
According to mother, patient had no complaints of paroxysmal noc-
turnal dyspnea, has known family DM II history, is a type II and
has diabetic retinopathy and macular edema with osteoarthritis.
```

Understanding the precise diagnostic terms to assign the correct set of ICD codes to the above EHR sentence involves the following steps:

- The detection of the correct ICDs: i) "E11.3219", which involves capturing every detail, such as the category of disease (type II diabetes), the anatomic site (diabetic retinopathy), the complication (macular edema), and the laterality (unspecified eye, by omission), ii) "M19.90" due to the mention of "osteoarthritis"

- Detecting the negation of "paroxysmal nocturnal dyspnea", and, therefore, not coding the record with the "R06.00" code

- Getting the relevant family history of the mention of "DM II" and comprehending that the record should not be coded with the corresponding "E11.9" code because it is related to a relative and not the patient.

As the example exhibits, understanding narrative, free text language to determine the correct set of ICDs conveyed in the EHR like an expert coder would is not a straightforward task.

1.2 Publications

We have published a total of 8 publications: 5 articles in 4 distinct JCR Q1 journals, 1 article in a JCR Q2 journal, and 2 international conference publications. These works are enumerated below together with the impact factor (IF) of the associated journals. Additionally, we was honoured with the CLEF eHealth 2020 international conference first prize.

1. **Expert Systems with Applications (2019)**
   Blanco A, Casillas A, Pérez A, Díaz de Ilarraza A. Multi-label clinical document classification: Impact of label-density. In Expert Systems with Applications. 2019 Dec 30; 138:112835.
   JCR: Q1 – IF: 5.452 – IF5: 5.448.

2. **Computer Methods and Programs in Biomedicine (2020)**
   Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. In Computer methods and programs in biomedicine. 2020; 188, 105264.
   JCR: Q1 – IF: 5.428 – IF5: 5.034.

3. **IEEE Journal of Biomedical and Health Informatics (2020)**
   Blanco, A., Pérez, A., Casillas, A., & Cobos, D. (2020). Extracting Cause of Death From Verbal Autopsy With Deep Learning Interpretable Methods. In IEEE Journal of Biomedical and Health Informatics. *2020; 25(4), 1315-1325.*
   JCR: Q1 – IF: 5.772 – IF5: 6.018.

4. **IEEE Access (2020)**
   Blanco, A., Pérez, A., & Casillas, A. Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time. In IEEE Access. *2020; 8, 183534-183545.*
   JCR: Q2 – IF: 3.367 – IF5: 3.671.

5. **CLEF eHealth (2020)**
   Blanco, A., Pérez, A., & Casillas, A. (2020). IXA-AAA at CLEF eHealth 2020 CodiEsp. Automatic Classification of Medical Records with Multi-label Classifiers and Similarity Match Coders. In CLEF Working Notes.
   GII-GRIN-SCIE: A- – SJR: Q – IF: 0.18
   This contribution was awarded the first position in the CodiEsp-D task.

6. **IEEE Journal of Biomedical and Health Informatics (2021)**
   Blanco, A., Pérez, A., & Casillas, A. Exploiting ICD Hierarchy for Classification of EHRs in Spanish through multi-task Transformers. In IEEE Journal of Biomedical and Health Informatics. *2021. 1374-1383.*
   JCR: Q1 – IF: 5.772 – IF5: 6.018

7. **Recent Advances in Natural Language Processing (2021)**
   Blanco, A., Remmer, S., Pérez, A., Dalianis, H., & Casillas, A. On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages with Fewer Resources than English. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*; (165–172).*
   GII-GRIN-SCIE: B – IF: 0.18
   International collaboration with Stockholm University.

8. **International Journal of Medical Informatics (2022)**
   Trigueros, O., Blanco, A., Lebeña, N., Casillas, A., & Pérez, A. Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. In International Journal of Medical Informatics, *157, 104615.*
   JCR: Q1 – IF: 4.046 – IF5: 4.768

In the following sections (1.2.1-1.2.5), we present a brief description of the five works that constitute the compilation for this thesis. The full publications can be found in Appendices A.1-A.5. Each work addresses research questions that arose to tackle the objectives listed in Section 1.1 related to each objective as summarised in Figure 1.4. Altogether, the main research question this thesis seeks to address is as follows.

Main research question: *How can we automatically assign the diagnostic terms enclosed in an EHR according to the ICD?*

Objective 1: Develop versatile text classifiers for diagnostic term classification.

- **RQ1** – Which classification methods can handle clinical document classification with lengthy EHRs?
- **RQ3** – Which is the most appropriate embeddings technique for clinical document classification?
- **RQ5** – Are the models versatile enough to be extensible to other medical specialities?

Objective 2: Perform extreme multi-label text classification on lengthy documents.

- **RQ4** – How do the input text and output labels characteristics affect the classifiers?
- **RQ6** – Is the transformer architecture robust for clinical document classification?

Objective 3: Predict explainable and coherent label sets of ICDs.

- **RQ2** – Can the consistency of the set of predicted labels be increased?
- **RQ7** – Can the hierarchical characteristics of the ICD improve the predictive ability of a model?
- **RQ8** – Can per-label attention aid with the prediction of coherent label sets?

Figure 1.4: This is a list of our research questions, grouped by objective and numbered according to their order of appearance in the publications.

Next (in Sections 1.2.1-1.2.5), we elaborate on each research question by introducing the articles included in the compendium. To present the research questions tackled in each work in an orderly manner, Figure 1.5 shows the article(s) in which each question is addressed. Figure 1.5 also depicts how each publication is related to each objective and research question.

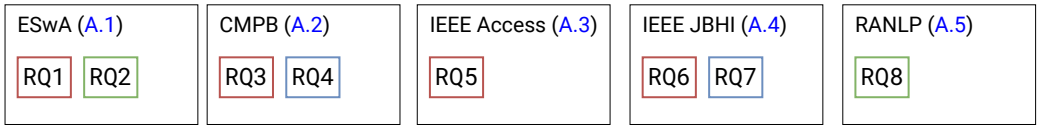


Figure 1.5: Overview of publications related to each research question and objective. The appendix where each article is available is listed in parenthesis.

1.2.1 Expert Systems with Applications (2019)

Reference

Blanco A, Casillas A, Pérez A, Díaz de Ilarraza A. Multi-label clinical document classification: Impact of label-density. In Expert Systems with Applications. 2019 Dec 30; 138:112835.
The full article is available in appendix A.1.

Abstract

This work aims to automatically designate the diagnoses enclosed in an electronic health record under the International Classification of Diseases. In natural language processing terms, a multi-label text classification task is associating each clinical record with a set of labels (i.e., ICD codes). We began our research with the artificial neural network methods, as neural approaches were the NLP trend, exhibiting superior performance to traditional methods (Li et al., 2018; Huang et al., 2019). Then, we stumbled upon an XMC problem and studied the impact of the label set size on its performance. We focused on the vast number of labels that characterise the ICD coding and proposed approximations to obtain reduced label sets based on the absolute and relative prevalence of labels across the EHRs along with consistency techniques for promoting the prediction of coherent label sets. We suggested transforming the label set into an approachable set by keeping labels that appear at least in 1% to 5% of EHRs.

Novelty

Studies in the field of clinical record multi-label classification had previously focused only on the binary relevance approach (Tsoumakas et al., 2009; Read et al., 2011): applying a binary classifier to determine the presence or absence of each ICD independently of the rest. Nonetheless, since medical conditions

are often correlated, independent classifiers were incapable of modelling relationships and could not ensure the consistency of the predicted label set. In this work, we intended to determine the extent to which binary classifiers limit model performance and whether the use of intrinsically multi-label classifiers increases consistency.

At the same time, far too little attention was paid to label sets with low-frequency codes (low-density label sets) in the antecedents (Gangavarapu et al., 2020). The scarceness of the labels made the prediction task even more troublesome. Although Gangavarapu et al. were working on one-shot or zero-shot approaches, research on systems that handle frequent and infrequent labels simultaneously was limited. Therefore, we explored the ability of different classifiers to deal with high- and low-density label sets.

Approach

To tackle the task, we investigated three neural network architectures and two strategies—binary relevance and multi-label output. We evaluated shallow and deep feed-forward network architectures as well as a recurrent model based on the bidirectional gated recurrent unit (GRU) architecture (Cho et al., 2014; Chung et al., 2014). We focused on models capable of capturing and modelling label dependencies on the output layer in addition to proposing label set consistency techniques. This work addressed the following research questions: i) **RQ1:** *Which classification methods can handle clinical document classification with lengthy EHRs?* ii) **RQ2:** *Can the consistency of the set of predicted labels be increased?*

Findings

RQ1 Prior studies had demonstrated that neural networks methods outperform traditional approaches. Consequently, dense features are a better choice than the sparse ones (Wang et al., 2016; Yeh et al., 2017; Baumel et al., 2017). Therefore, we started our research with neural methods, exploring a variety of model architectures including i) shallow with a non-linear binary logistic regressor (BLR), ii) deep with a deep neural network (DNN), and iii) recurrent neural networks (RNN) with a bidirectional recurrent neural network with GRU units (BiGRU). In this work, we found that the recurrent models overcome the non-recurrent ones, especially the bidirectional recurrent neural network.

Additionally, we found that the appropriate features for neural methods are the dense features. We obtained F1-scores for the DNN and the BiGRU of 72.3 and 79.7 for the MIMIC and 58.8 and 72.5 for the Osakidetza dataset, respectively. Thus, we confirmed that word embeddings with recurrent models (BiGRU) are a more appropriate option for EHR classification in Spanish than document embeddings with non-recurrent classifiers (BLR, DNN).

RQ2 The non-linear binary logistic regressor baseline implemented for this work could not capture and model the relationships among labels. On the contrary, the deep and recurrent neural networks are intrinsically multi-label classifiers, modelling the relationship among labels on the output layer. The improvement of the DNN and RNN with respect to the baseline showed that the relationships among labels are moderately captured. Additionally, we developed techniques for diminishing the prediction of incompatible label sets and promoting the prediction of related labels. These label-consistency rectification methods slightly improved the results on the Osakidetza corpus while imposing no extra cost during the training stage and negligible post-processing cost. Specifically, we kept a class weighting procedure to consider skewed class distributions and a threshold selection criterion to prevent the bias towards frequent classes for future works.

Concluding Remarks

The recurrent neural networks exhibited superior performance. However, the characterisation of the clinical documents had room for improvement. The most promising research idea was exploring and determining which word embeddings worked best with the recurrent networks on Spanish EHRs. Another critical point was understanding how the different input and output characteristics affect the performance of the classifiers. Despite the promising results, we believe that further research should deepen on the label set consistency rectification approaches because they could leverage the accuracy attained by the inference algorithms. Further improving relationship capture and modelling remains an open question. We suggested strategies such as statistically driven approaches (e.g., correlation analysis) (Zhang and Schneider, 2011) or the incorporation of the hierarchical structure of the ICD.

1.2.2 Computer Methods and Programs in Biomedicine (2020)

Reference

Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. In Computer methods and programs in biomedicine. 2020; 188, 105264.
The full article is available in appendix A.2.

Abstract

In this work, we continue tackling the automatic classification of EHRs according to the ICD. We applied a larger dataset than the used in our previous work (Blanco et al., 2019), which was also obtained from the Basque Country Health System. We focused on transforming the input and output, exploring different input texts and label sets. We continue our research from the recurrent neural network methods, which were established as the state-of-the-art in NLP (Tang et al., 2015; Zhou et al., 2016) while also exhibiting superior performance in our past work (Blanco et al., 2019).

Novelty

Although electronic health records are unstructured text reports, they are usually made up of distinguishable segments which can be extracted using patterns or regular expressions in some datasets. However, extracting the different parts from the EHRs is not a trivial task (Goenaga et al., 2021). Another way to address the task is as multi-class classification, finding entities and mapping those entities to ICD codes. For example, Suominen et al. (2018) only addressed the mapping task. It is a more accessible approximation because the aim is only to classify the standard terms, and input is less than 10 words long on average versus the hundreds of words of the complete EHRs. One concern that was not yet clear was the impact of the part of the clinical record on the performance of the classifiers (Berndorfer and Henriksson, 2017; Gangavarapu et al., 2020). We investigated the utility of using a particular section versus the full health record for ICD classification. We started by focusing on the "diagnostic impression" section, where most ICDs are localised. However, while it is true that much information is held within

that section of the EHR, we found that there is also valuable information in other sections.

In addition, no research was found surveying the influence of label granularity for ICD classification. Apart from the label set cardinality, the degree of specificity of the output labels can impact the performance because the discrimination difficulty varies. We attempted to derive several sets of labels from the fully specified ICD coded labels at different levels of granularity and assess their impact on the prediction ability of the classifier in relation to the number of labels.

Approach

We investigated four RNN architectures with variations on the embedding layer while applying a non-recurrent model as the baseline to address this work. Current models could capture and model the label dependencies to some extent with the sigmoid output layer. Thus, we kept most of the architecture of the previous models and the best-performing consistency techniques from our previous work (i.e., the class weighting and threshold selection criterion) (Blanco et al., 2019). Then, we concentrated on the text representation layer and tested three variants, including standard, meta, and contextual embeddings. The experiments explored a broad set of label sets while varying cardinality and granularity. This work addressed the following research questions: i) **RQ3:** *Which is the most appropriate embeddings technique for clinical document classification?* ii) **RQ4:** *How do the input text and output labels characteristics affect the classifiers?*

Findings

RQ3 We found substantial differences among the different embeddings approximations, but contextual word embeddings were the most solid choice in terms of F1-Score, concretely, the ELMo (Embeddings from Language Model) embeddings. The F1-score achieved with the regular word embeddings was 49.8, while the performance increased to 54.3 with the ELMo embeddings. Nonetheless, it is worth mentioning that the meta-embeddings showed strengthened performance (53.9 F1-score points) compared to the regular word embeddings applied for the construction of the meta-embeddings. Note that the results are not directly comparable to those from our previous work because the

corpus and label sets differ. We delve deeper into this matter in the discussion in Section 1.3.

RQ4 The label set cardinality heavily impacts the ability of classifiers to perform multi-label classification on ICD codes. We tested to what extent each neural architecture degrades its performance as the label set density decreases and the size increases. We found that it is possible to increase the performance of the classifiers by working with a higher granularity without reducing the number of labels. The finding was depicted in an experiment with two identical runs except for the label set: one with fully-specified labels (i.e., the lowest level of granularity) and the other adding more labels but with higher granularity (i.e., block labels). We achieved the best performance with the higher granularity label set despite having more labels. This result could be explained because a lower granularity implies a higher degree of label specificity. Therefore, the discrimination between labels is a more complex task.

Contrary to our expectations, this study found that employing full documents leads to better results than focusing only on the "diagnostic impression" section. The results were 54.3 F1-score points when feeding only the diagnostic section and 63.2 when supplying the whole document to the classifier. This could be because, throughout the EHR, there are also explicit and implicit mentions of medical conditions related to ICD labels. Additionally, recurrent models are specialised in processing long sequences of texts.

Concluding Remarks

Our main findings were that the optimal option was to apply ELMo contextual embeddings while feeding the whole document to the classifier. A better understanding of which classification approach was the most appropriate must still be considered because the field was constantly evolving, and there were other approaches to contextual embeddings that were interesting. In example, approaches based on the transformer architecture, like BERT (Devlin et al., 2018). Furthermore, a study on how the prediction could be explained to practitioners was also needed. Other opened research questions include detecting different sections (i.e., patient history, prescriptions), feeding these sections separately to the model, or even leveraging structured text reports.

1.2.3 IEEE Access (2020)

Reference

Blanco, A., Pérez, A., & Casillas, A. Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time. In IEEE Access. *2020; 8, 183534-183545.*
The full article is available in appendix A.3.

Abstract

In our previous works (Blanco et al., 2019, 2020d), we processed the EHRs with RNNs and different embedding approximations (e.g., standard-, meta-, and contextual embeddings). Here, we tested models from past work on their versatility and ability to generalise in two areas: across time (i.e., as time moves forward), and across various hospital services or medical specialities. The experiments were designed to test if the models are robust enough or if their performance degrades when predicting future data or EHRs from different specialties than those used to train the model. Regarding the methods, we took the best classifier from our previous work (Blanco et al., 2020d)—the BiGRU model with ELMo embeddings—and evaluated its behaviour when facing changes in the training and testing data. To investigate and analyse the adaptability of the models, we examine their resilience over data from two years and six hospital services from two different hospitals. Additionally, we considered the categorisation performance while estimating ICD codes with varying degrees of granularity.

Novelty

Some antecedents reduced the label set size based on the frequency of the labels to disregard those ICDs with low support (Névéol et al., 2018; Gangavarapu et al., 2020). While it is a practical approach to get more reasonable label sets, evaluating the models with the complete set of labels is also interesting. In this work, we tested both approaches. First, we performed experiments with the unrestricted label set, predicting all the labels (6,918 labels on the MIMIC and 2,554 on the Osakidetza dataset) available in the data. Further, we conducted experiments with two reduced label sets, restricting the label set by keeping only the ICD labels that appeared in at least ∼1% and ∼5% of the documents, respectively.

Another concern is that models are not usually tested across time. A thorough search of the relevant literature yielded no data on how the models would behave after a given time has elapsed since deploying them to a production environment (i.e., a hospital). Therefore, we evaluated the resilience of the system over time. The change in data across time could be critical, as hospital personnel's EHR writing or coding styles may evolve as time passes. Then, could the system keep performance constantly, need to be retrained at given time intervals, or require continuous training? The purpose is to determine to what extent a predictive model drawn from historical data can forecast EHRs in future years.

We designed an experiment with non-overlapping data from two consecutive years to test the hypothesis. We tested the model trained with data from the first year to forecast EHRs from the subsequent year. After that, we compared the results with the performance obtained when training the model with all the available data. In other words, we studied the penalty in performance suffered from not keeping the models updated. We intended to determine the most effective strategy for continuous training (i.e., quantify the loss when retraining the models in given time frames versus a continuous training approach).

Similarly, little research had been done regarding the adaptation of the models to different hospitals and medical specialties (Pérez et al., 2018). Hence, we also evaluated the system across medical services from two different hospitals. One of our concerns was the scarcity of data. We wondered if a general system trained on EHRs from discharge summaries from multiple hospital services (e.g., cardiology, nephrology, psychiatry, and others) can combine and accurately capture syntax and semantic nuances from each speciality. The other possibility was that accurately encoding EHRs from a single service would inevitably require the system to be trained on EHRs solely from that service, as EHRs from other services may convey an unmanageable quantity of irrelevant vocabulary, resulting in a distorted outcome.

Approach

To this end, we apply our multi-label classification model based on a bidirectional recurrent neural network with GRU units, utilising the ELMo embeddings, according to our findings from previous works (Blanco et al., 2020d). We drew the following main research question: i) **RQ5:** *Are the models versatile enough to be extensible to other medical specialities?*

Findings

RQ5 The general trend is to train generalist models, namely, models trained on EHRs from various hospital services (Blanco et al., 2020d). However, the medical speciality has an important impact on model performance. Our results indicated that models trained by speciality performed better than generalist models when trained on specialty label sets. In four out of the six specialties, the results improved in terms of F-Score when the models were trained on EHRs from a specific medical specialty. The mean improvement obtained with the models trained by speciality was about 14 points. There were some notable increases, such as the 30 points improvement (from 21.28 to 52.38) in the digestive speciality. On the two medical specialities that performed better with the generalist model, the gain was only around 3 and 6 points. Admittedly, training speciality models implies an additional expense; it is necessary to train several models for each medical service and to be limited to more restricted speciality-related label sets.

Concluding Remarks

Interestingly, the relationship between the increase in the label set size and the reduction in performance is not exponential, as could be expected due to the exponential growth of the label relationships. Indeed, the results indicate a smaller performance drop than expected. Nonetheless, although the models are strong enough to accurately categorise specific EHRs from future years when trained purely on data from previous years, there is a performance penalty. Adding more years (i.e., more EHRs) to the training set improves performance.

As NLP technology advances, new architectures arise. Therefore, its performance in clinical text mining and multi-label classification must be tested. Specifically, language model (LM) classifiers are promising. Additionally, if the novel architecture delivers sound results on the task, new research questions will arise. In summary, instances of transformers and other language model classifiers must be developed to solve the ICD classification task. If the new architecture is successful, we feel that a vital issue for clinical multi-label classification would be the leverage of unsupervised, in-domain, and more similar data to further pre-train the LMs before applying them to the downstream task such as the multi-label classification of health records.

1.2.4 IEEE Journal of Biomedical and Health Informatics (2021)

Reference

Blanco, A., Pérez, A., & Casillas, A. Exploiting ICD Hierarchy for Classification of EHRs in Spanish through multi-task Transformers. In IEEE Journal of Biomedical and Health Informatics. *2021. 1374-1383.*
The full article is available in appendix A.4.

Abstract

The transformer architecture is state-of-the-art in NLP, bringing improved contextual awareness and mitigating catastrophic forgetting from RNNs. In this work, we test the transformer architecture for multi-label classification, developing a multi-label classifier on top of a BERT model, adapted to handle the ICD as the label set. Since we employ EHRs written in Spanish, the primary problem is the scarcity of resources compared to English, which we attempt to lessen through the language modelling and sub-word encoding features of the transformers. Furthermore, we are concerned about consistency in labelling. Thus, we conceive a technique to tackle the relationship among labels on the classification head on top of the language model of the BERT model to gain consistency in co-morbidity prediction.

Novelty

Past work had explored the impact of pre-training and fine-tuning LMs using in-domain vast publicly available datasets Gu et al. (2020); Zhang et al. (2020). However, much less was known about the effect of the further fine-tuning with smaller but closely related datasets. We proposed a strategy that benefited from the most similar possible unsupervised EHRs to the EHRs to be classified (i.e., unsupervised EHRs that come from the same hospital). In the field of automatic clinical coding, it is common to have just a tiny share of the available EHRs manually coded. The rest of the dataset samples, which are not coded, are discarded. We propose to use the usually disregarded unsupervised (i.e., not coded) EHRs for further pre-training of the LM.

Another concern is modelling the relationship among labels with hierarchical classification models. Admittedly, there is extensive research on hierarchical models, but most ignore the possibility of leveraging the ICD

hierarchy (Rios and Kavuluru, 2018). On the other hand, previous research had established that multi-task architectures improve the capacities of language models in diverse NLP tasks (Yang and Shang, 2019). Nevertheless, most studies had focused on tackling related but different tasks such as several classification tasks (i.e., part-of-speech and named entity recognition), or even combined image and text classification—in summary, tasks with a direct dependency/structural relationship. The limitation of these approaches for clinical tasks is that they require more than one label set related to each health record to leverage the multi-task learning paradigm. However, obtaining thousands of labelled records is a time-intensive and expensive task. To bring both questions together, we proposed to exploit the hierarchical nature of the ICD to extract the required label sets to allow multi-task learning as well as benefit from the shared information from the different hierarchical levels.

Approach

We implemented a hierarchical head for multi-label classification, benefiting from the ICD hierarchy through multi-task classification. Regarding the LMs, we explored two BERT models (generalist versus in-domain pre-trained models) with BERT Multilingual and BioBERT (Lee et al., 2020). This work addressed the following research questions: i) **RQ6:** *Is the transformer architecture robust for clinical document classification?* ii) **RQ7:** *Can the hierarchical characteristics of the ICD improve the predictive ability of a model?*

Findings

RQ6 In this work, we switched from the BiGRU model from (Blanco et al., 2020d) to a new classifier based on transformers. As expected, even with the most simplistic form of a multi-label classifier built on top of a BERT multilingual model, it outperformed the RNN. The F1-score obtained with the RNN on (Blanco et al., 2020d) was 39.9, while the BERT model achieved 46.4 on the most similar label set.Given the results, we focused on further improving and adapting the transformer-based models to the ICD classification tasks; however, the architecture imposed some significant limitations. First, most transformer models have a tight limitation on the maximum number of tokens that they

can handle per sample. For example, regular BERT models have a limitation of 512 tokens—far less than the average number of words of the EHRs from our datasets (around 800 for Osakidetza and 1,400 for MIMIC). Furthermore, extra work was required to get meaningful per-word attention that was comparable to attention outputs of the RNN models.

RQ7 We delved into the capture and modelling of the relationships among labels, hypothesising that the hierarchy of the ICD could be applied to improve the consistency of the predicted labels. We knew that transformer models benefited from the multi-task setting, so our proposal combined the two aspects. We implemented a hierarchical head for multi-label, multi-task classification, allowing us to tackle three classification tasks simultaneously. The multi-task learning paradigm distributes pertinent information across related tasks to enhance each individual task. In this work, we applied a three-level hierarchy based on the granularity of the labels. We found that the multi-task improved the performance; the best non-hierarchical approach got an F1-score of 46.4, while the hierarchical approach with the same settings achieved 50.8. We improved the ICD classification performance by 9% without requiring newly supervised (i.e., annotated) data, capitalising on the synergies between the various sets of labels derived from the ICD hierarchy.

Concluding Remarks

An exciting way to continue this study was to investigate additional models based on the transformer architecture apart from BERT. Each model has its own unique features; finding the one that better fits the task is not trivial and could lead to improved results. The attention mechanism on multi-label classification models is crucial. It serves both to improve the results and enable interpretability capabilities. Therefore, extending, improving, or developing alternative attention mechanisms to the self-attention mechanism from BERT is another area requiring further research. Finally, the multilingual language models, following the idea of BERT Multilingual, could be of interest to the languages with less biomedical NLP resources than English—they could benefit from the available biomedical data from several languages.

1.2.5 Recent Advances in Natural Language Processing (2021)

Reference

Blanco, A., Remmer, S., Pérez, A., Dalianis, H., & Casillas, A. On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages with Fewer Resources than English. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*; (165–172).*
GII-GRIN-SCIE: B – SJR: Q – IF: 0.18
International collaboration with Stockholm University.
The full article is available in appendix A.1.

Abstract

From our previous work (Blanco et al., 2021a), we learned that transformer models were appropriate for ICD multi-label classification. In this work, we researched new architectural modifications (i.e., a new per-label attention mechanism) while tackling a new challenge: developing a single model to code EHRs in English and two additional languages with fewer biomedical resources, Spanish and Swedish. Another challenge was to achieve a comparable label set to promote comparability, which was challenging because the data came from three different hospitals in three countries (Spain, Sweden and the United States of America). The main goal was to develop a model that was not language-bound to solve the classification task on the three languages with the same model. We developed a solution with a shared language model for the EHR representation and a classification head explicitly trained for each language for the multi-label classification. The head was the only module trained from scratch for each language. Furthermore, we aimed to extend the attention mechanism, building an additional attention module. Hence, we implemented a language-agnostic model with a novel attention module, extending the standard self-attention mechanism with per-label attention. This module also served as an interpretability system; one benefit of our per-label attention was that a specific attention weight was computed for each word-label pair.

Novelty

In this work, we focused on a single medical speciality, Diseases from the Gastrointestinal System. We obtained a comparable label set with 157 ICD codes in the three applied datasets (Spanish, Swedish, and English). Note that it is easier to differentiate conceptually distinct illnesses (e.g., gastrointestinal versus cardio-pulmonary ailments) than to discriminate between two medical conditions within the medical speciality. Therefore, focusing on the classification of similar diagnostic terms further motivated the development and application of the per-label attention mechanism.

It is typical to have models tied to a given language, which could be used for developing specialised models for languages with a considerable amount of resources, such as English. Nevertheless, it may not be the best option for low-resource languages (such as Swedish) since data availability may be insufficient for training the models only with the primary language (Pires et al., 2019). Hence, it is beneficial to leverage multilingualism. To that end, we proposed a classifier with a shared language model among the three considered languages. In other words, we benefited from the synergistic impact of a multilingual approach in which the three combined languages (including English) gain from a more comprehensive volume of data.

A growing body of literature recognises the usefulness of attention mechanisms to improve classification performance (Mullenbach et al., 2018; Ji et al., 2021) not only for transformers, but also for recurrent neural networks and other architectures. To that end, our attention mechanism built on the multi-label classification head enabled the model to assign a weight to each word separately for each label.

Approach

To take these ideas into practice, we continued working with the BERT Multilingual model to build our multi-label classifier. However, the model incorporated the attention mechanism for computing specific attention weight for each input token and label pair. This work addressed the following research questions: i) **RQ8:** *Can per-label attention aid with the prediction of coherent label sets?*

Findings

RQ8 In multi-label classification, a given word can be relevant for indicating the presence or absence of a specific label while being irrelevant for others. Thus, per-label attention (i.e., computing how relevant is the presence or absence of a given word for each label separately) is required.

Experiments demonstrated that the model with per-label attention was superior to the standard BERT model for ICD categorisation, producing similar results for the three different languages. When the confusion matrices were analysed, we deduced the source of improvement for the per-ICD model; although the true negatives remained close to those of the other model (since with a large label set, the majority of classes are negative), the number of true positives increased by about 100%. Similarly, the false negatives were reduced by around 20%. We observed this pattern in every language. Moreover, the per-ICD model also exceeds the standard BERT model in interpretability since it can export the attention weights for its visualisation. We concluded that per-label attention improves the capacity for discriminating between semantically related labels.

Concluding Remarks

A weakness was that an EHR could be longer than the maximum sequence length of BERT models. Although we already demonstrated that the whole document contains valuable information (Blanco et al., 2020d), a reasonable question was whether to feed the entire EHR or to feed only the most informative parts. Further research may also involve applying BERT models that have been trained for particular languages, such as the BETO model for Spanish (Cañete et al., 2020) or the KB-BERT model for Swedish (Malmsten et al., 2020). Also, further research could include the ability to learn continuously by taking advantage of human expert knowledge through a human-in-the-loop feedback system incorporated into a visualization tool. Nonetheless, per-label attention helps in modelling the relationships between ICDs. If the same term is associated with two distinct ICDs, it suggests that both are related. Hence, the BERT model with per-label achieves the best results, increasing the F-Score by 34.9, 34.5, and 5.93 points for the Spanish, Swedish and English datasets, respectively, compared to the standard BERT model.

1.3 Discussion

This final section contextualises the results and discusses the best scores obtained and the progression of the developed classifiers for the automatic coding of ICDs. Comparing studies within the field of multi-label clinical classification is notably difficult due to the diversity and confidentiality of most datasets as well as the absence of standard datasets and benchmarks (Gu et al., 2020). Moreover, even in works which apply reference datasets such as the MIMIC-III, authors mine the datasets differently. As a result, there are large variations in the input text and output label sets and, therefore, in the results (Sammani et al., 2021).

In our work, the variation of experimental setups throughout the publications causes the results to be heterogeneous and, thus, challenging to compare. The reasons are multiple, but mainly, the expansion of the datasets by acquiring new data and the transformations of the datasets required while exploring new techniques. Note that developing more capable approaches enabled us to progressively increase the difficulty of the task (i.e., by expanding the label set size or by addressing more specific codes). The best result among our works (A.1-A.5) is obtained by the multi-label classifier with BERT Multilingual as the language model and per-label attention from A.5. The results are 41.1, 40.65, and 38.36 in terms of weighted F-Score for Spanish, English, and Swedish, respectively. These results are hardly comparable with previous works since the applied corpus, label sets, label granularity, and number of labels differ due to experimental setup constraints.

Accordingly, to offer a more straightforward overview of the evolution of the developed approaches, we took the largest Osakidetza dataset, described in Table 1.1, and extracted six different subsets varying the label set size and granularity of labels. Then, we applied the best model from each work to each subset to properly compare the performance. Specifically, we conducted the experiments on the three levels of granularity depicted in Figure 1.1 (full-code, block and chapter). For each granularity level, we tested two label set sizes: i) The non-reduced label set (i.e., maintaining all the original labels) referred to as Osa-0r, and ii) a reduced label set, preserving solely the labels that appear in at least 1% of the EHRs, referred to as Osa-1r. The results obtained with the best model developed for each work (A.1-A.5) are shown in Figure 1.6. The size of the label set for each experiment is denoted by N in the legend of each figure.

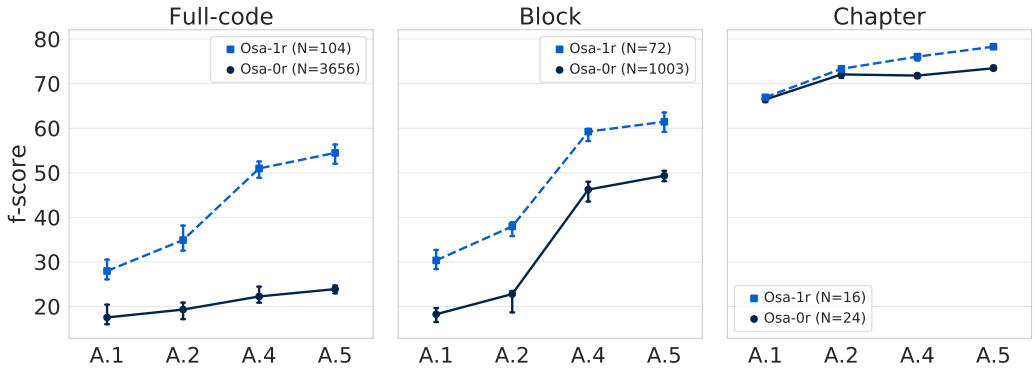


Figure 1.6: Experimental results of the best models from the publications, namely: A.1) BiGRU with standard embeddings, A.2) BiGRU with ELMo contextual embeddings, A.4) BERT Multilingual with hierarchical multi-task architecture, and A.5) BERT Multilingual with per-label attention mechanism. Experiments were conducted on the Osakidetza dataset at different granularity levels (full-code, block, chapter) and varying label set sizes (N indicates the number of labels). The model from A.3 is excluded because there were no architectural changes compared to the model from A.2.

Figure 1.6 shows the aggregated results from the five runs for each experiment in weighted F-Score, with the median being the estimate of central tendency with confidence intervals (95% CI). The obtained results reveal the upward trend of the classification performance attributable to the advancement of the architectures (i.e., from RNNs to transformers) and the incorporation of more specialised adaptations (i.e., the hierarchical multi-task architecture or the per-label attention mechanism). Regarding the results from the A.4 and A.5 models, the performance of the model from A.5 is just marginally better for non-reduced labels (Osa-0r) than those obtained with the A.4 model. The reason is that the model from A.5 does not incorporate the hierarchical multi-task architecture from the A.4 model. Thus, it is reasonable to expect that there is room for improvement by combining both strategies.

The results show the outcome of adapting the different deep learning architectures to the ICD coding task. We confirmed the strengths and weaknesses of each approach and implemented specific techniques to deal with the nuances of medical coding. Each approach has its unique features that translate to individual qualitative characteristics. To illustrate this point,

consider the following examples from EHRs.

- **Discriminate meaning based on context:** The BiGRU with ELMo contextual embeddings (A.2) had an increased capacity in the representation of words owing to the contextual embeddings.

  For example, this model could differentiate the meanings that the word, "cold", can have in the medical jargon depending on its context. "Common cold" is a disease, "cold temperature" indicates a symptom, and "COLD" is the acronym of Chronic Obstructive Lung Disease.

- **Multilingual ability:** The BERT Multilingual model based on the transformer architecture (A.4 and A.5) achieved stronger results. The main reason for this is that its text representation capacity is superior, partly due to the sub-word encoding required for the multilingual ability.

  For example, many medical terms have shared roots (i.e., "insulin", "insulinoma") and even similar word formations in several languages (i.e., "insulin" in English and "insulina" in Spanish). This fact allows the transfer of learning from multiple languages.

- **Hierarchical multi-task architecture:** The BERT Multilingual model with a multi-task hierarchical architecture (A.4) benefits from less detailed predictions of higher granularity codes to predict more specific codes.

  For example, the model that accurately predicts that a given EHR addresses metabolic diseases ("E" label, at the chapter granularity level) more easily predicts more specific codes within the hierarchy such as the fully-specified code "E11.9 – Type 2 diabetes mellitus."

- **Attention mechanisms:** The BERT Multilingual model with the per-label attention (A.5) could compute the significance of each word for each ICD code. This mechanism internally models the relationship between labels, improving its precision and offering improved result interpretability.

  For example, a particular word may be relevant for two different ICDs (i.e., the word "alcohol" is relevant for labels "F10.19 – Alcohol abuse" and "K76.0 – Fatty liver"). This information suggests that both labels are related; the model can benefit from that knowledge.

In conclusion, we have developed an end-to-end prototype to classify EHRs with multiple models and approaches. The software prototype includes the transformation and pre-processing of source data, numerous deep learning classification approaches, post-processing, evaluation of results, and visualisation of predictions. The text processing procedures do not heavily depend on word-form alterations such as removing characters, tokenisation, or lemmatisation. Most text pre-processing is limited to transforming the input sources to a standard input structure. This pre-processing approach yields more versatile models that are not tied to specific prerequisites (i.e., particular languages or medical specialities). Avoiding heavy text preprocessing techniques prevents ad-hoc solutions that limit the adaptability of the models and their deployment to real-world scenarios. Altogether, each approach improved the previous one. Addressing every research question, we focused on particular challenges and found ways to cope with them, yielding subsequent improvements in terms of predictive ability.

2 Conclusions

This concluding chapter reflects on the lessons learned, conclusions reached, and unexplored areas for further research. First, we provide a summary of the research in Section 2.1, aggregating all our findings and conclusions throughout the publications as well as providing a list of the contributions. Finally, in Section 2.2, we delve into the most promising future work and the lines of research that remain open.

2.1 Concluding remarks

This thesis was undertaken to explore the automatic multi-label classification of EHRs according to the International Classification of Diseases. The task was to solve a multi-label text classification problem: a supervised learning task founded on categorising text documents with a set of non-mutually exclusive labels. The main **research question** was: *how can we automatically assign the diagnostic terms enclosed in an EHR according to the ICD?*

Since medical coding is a particular text classification task in which an understanding the natural language involved is crucial, we raised the following hypothesis: NLP and deep learning classification techniques can mimic the process that expert human coders accomplish. To determine if this was true, we established the **main objective** of the thesis: *develop a method to automatically determine the diagnostic terms enclosed in an EHR according to the ICD*. Three sub-objectives were derived from it. The assessment of the expected outcomes and a summary of our conclusions from the publications follows.

- **Objective 1** *Develop versatile text classifiers for diagnostic term classification*: We decided to begin our research journey directly with deep learning models. We started from simple models, but continuously developed, adapted, and conceived new techniques to keep innovating on the state-of-the-art NLP models. We confirmed that, as with general domain classifiers, recurrent neural networks and transformer models are the most appropriate choices for clinical text multi-label classification. We concluded that, while it seems that transformer models are superior to RNNs, the latter still have advantages worth considering for this task such as the capacity to handle longer sequences.

  We found that the diversity of health records and their intrinsic characteristics are among the most significant challenges in medical coding. We deal with an extensive vocabulary, specialised medical jargon, as well as an abundance of abbreviations, acronyms, and spelling mistakes. Moreover, models must adapt to EHRs in various languages or from different medical specialities and hospitals. However, deep learning techniques are still the most suitable option due to their multilingual capabilities and capacity to work with sub-word information.

  Regarding text representation, we confirmed that contextual word embeddings conveyed with language models based on either RNNs or transformers are the best options for clinical text classification.

  To accomplish the objective, we developed models that could handle either fragments of EHRs or the whole documents as input (Blanco et al., 2020d). The models were adaptable to datasets from several hospitals and medical specialities (Blanco et al., 2020a) in addition to different languages (Blanco et al., 2021b).

- **Objective 2** *Perform extreme multi-label text classification on lengthy documents*: The length of the input text significantly impacts the behaviour and performance of the models. A fundamental finding was that, with models that process the text as sequences (i.e., RNNs), the best option was to feed the full health record instead of fragments such as the diagnostic section. This was true because there is meaningful information throughout the entire document (Blanco et al., 2020d). This conclusion is valuable for the previously mentioned aim of limiting the text pre-processing to avoid ad-hoc solutions. It is desirable to feed the whole document as raw text and avoid any pre-processing, such as

extracting sections or text fragments from the EHRs. The methods for extracting sections are usually tied to a given language. Additionally, most hospitals have their own formats for their medical records; there is a lack of predetermined structure of the health records.

The differing granularity levels of the ICD unlock exciting possibilities to tackle with the classification models. We showed that hierarchical and multi-task classifiers could improve performance on extreme label sets benefiting from several granularity levels from the ICD hierarchy (Blanco et al., 2021a). This is also supported by the results of Section 1.3. Nevertheless, labelling EHRs with coarse-grained labels might have utility apart from improving the fully-specified code predictions. For example, labelling with the ICD chapter granularity level might be efficient for clinical documentation.

- **Objective 3** *Predict explainable and coherent label sets of ICDs*: The relationship among labels is relevant in any multi-label text classification task. However, it becomes crucial when the label sets are highly and complexly correlated, as in the case of relationships among medical conditions (Yao et al., 2017; Lee et al., 2018; Cheng et al., 2020). For that reason, we opted for models that intrinsically modelled the multi-label task, taking into account the label inter-relationships. However, applying the common fully connected deep learning layers was insufficient to capture the relations among thousands of labels. We have explored other techniques that work jointly with the intrinsically multi-label layers and found compelling possibilities.

  According to our experiments, attention mechanisms seem to be powerful methods for handling the relationship among labels and improving the classification results. This is especially true for per-label attention mechanisms operating with high cardinality label sets (Blanco et al., 2021b). This claim is also corroborated by the results presented in Section 1.3. Moreover, exploiting the hierarchical nature of the ICD through the multi-task setting with the different hierarchical levels is a strong choice (Blanco et al., 2021a). Apart from the performance boost achieved without the need for additional supervised data, the model may be trained on several tasks concurrently with little additional computational work.

The findings presented in this dissertation contribute to a better understanding of the ability of current machine learning techniques to conduct medical classification in a similar way to expert human coders. Overall, the thesis concludes that the models based on the transformer architecture—benefiting from the ICD hierarchy through a multi-task architecture and equipped with per-label attention—are the most appropriate choices. Answering the main research question and achieving the objectives throughout the thesis yielded several **contributions** apart from the publications. These contributions comprise the software modules developed to tackle the multi-label classification task, library extensions, and compatibility with tools for model interpretability.

The complete software packages developed for each work were released with the publication of the articles: i) The data analysis, preprocessing, and evaluation modules, along with the classification models are available in (Blanco et al., 2019). ii) The preprocessing, training and inference utilities for running the BiGRU with ELMo contextual embeddings models are available in (Blanco et al., 2020d). iii) The single- and multi-task models that extend the Hugging Face Transformers library (Wolf et al., 2020) with NeAt-Vision (Baziotis, 2018) integration for visualising the per-label attention is available in (Blanco et al., 2021b).

2.2 Future work

Clinical multi-label text classification is a field that has experienced a breakthrough in recent years—partly due to the significant development of NLP thanks to deep learning. However, progress has been slowed compared to other sectors because DL methods require large amounts of data; medical data is sensitive and complex to obtain. In our view, the most promising areas of research for the near future are summarised here.

- **Named entity recognition**—Applying medical named entity recognition techniques as a pre-processing step: The extracted medical entities could be used as features for the classification models. The detection of the medical entities could help the model recognise relevant information such as disorders and their characteristics like laterality, severity, or body-part or negation cues (Santiso et al., 2020).

- **Increase data availability**—Text augmentation techniques with unsupervised electronic health records: We discovered that obtaining unannotated data to improve the language model is critical for languages with scarce biomedical NLP resources. However, there is a bottleneck for obtaining EHRs in languages other than English (Névéol et al., 2018; Dalianis, 2018). Unsupervised automated translation (Artetxe et al., 2017) of EHRs from languages with more resources (i.e., English) into other languages could result in the production of hundreds of thousands of EHRs. Similarly, abstractive summarisers (Savelieva et al., 2020) may be used to construct alternate versions of current EHRs, thereby multiplying the number of available health records for training the models.

- **Improve the ability to handle lengthy documents**—Long-range transformers: We propose using long range transformers to analyse lengthy documents. Thanks to modifications in the self-attention mechanism, transformer models, such as the Longformer (Beltagy et al., 2019) or BigBird (Zaheer et al., 2020), increased the maximum sequence length above the 512 tokens of the BERT model up to 4096 tokens (Tay et al., 2021). This characteristic is convenient for working with EHRs containing thousands of words.

List of Abbreviations

AI Artificial Intelligence.

BERT Bidirectional Encoder Representations from Transformers.

BiGRU Bidirectional Recurrent Neural Network with GRU units.

BLR Binary Logistic Regressor.

CNEAI National Assessment Committee for Research Activities.

DL Deep Learning.

DNN Deep Neural Network.

EHR Electronic Health Record.

ELMo Embeddings from Language Model.

GRU Gated Recurrent Units.

ICD International Classification of Diseases.

IF Impact Factor.

JCR Journal Citation Report.

LM Language Model.

ML Machine Learning.

NLP Natural Language Processing.

NLU Natural Language Understanding.

RNN Recurrent Neural Network.

RQ Research Question.

XMT Extreme Multi-label Classification.

XMTC Extreme Multi-label Text Classification.

Bibliography

Araujo, L., Martinez-Romo, J., Turmo Borras, J., Padró, L., Casillas Rubio, A., and Gojenola Galletebeitia, K. (2021). DOTT-HEALTH: Desarrollo de tecnología aplicada a textos para el soporte de diagnosis, prevención y gestión de instituciones de salud. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021): co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021): Málaga, Spain, September, 2021*, pages 13–16. CEUR-WS.org.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised Neural Machine Translation. *CoRR*, abs/1710.11041.

Baumel, T., Nassour-Kassis, J., Elhadad, M., and Elhadad, N. (2017). Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment. *CoRR*, abs/1709.09587.

Baziotis, C. (2018). Neat (Neural Attention) Vision. https://github.com/cbaziotis/neat-vision.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China,*

November 3-7, 2019, pages 3613–3618. Association for Computational Linguistics.

Berndorfer, S. and Henriksson, A. (2017). Automated Diagnosis Coding with Combined Text Representations. *Studies in health technology and informatics*, 235:201–205.

Blanco, A., Casillas, A., Pérez, A., and Díaz de Ilarraza, A. (2019). Multi-label clinical document classification: impact of label-density. *Expert Systems with Applications*, 138:112835.

Blanco, A., Pérez, A., and Casillas, A. (2020a). Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time. *IEEE Access*, 8:183534–183545.

Blanco, A., Pérez, A., and Casillas, A. (2020b). IXA-AAA at CLEF eHealth 2020 CodiEsp. Automatic Classification of Medical Records with Multi-label Classifiers and Similarity Match Coders. In Cappellato, L., Eickhoff, C., Ferro, N., and Névéol, A., editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Blanco, A., Pérez, A., and Casillas, A. (2021a). Exploiting ICD Hierarchy for Classification of EHRs in Spanish Through Multi-Task Transformers. *IEEE J. Biomed. Health Informatics*, 26(3):1374–1383.

Blanco, A., Pérez, A., Casillas, A., and Cobos, D. (2020c). Extracting Cause of Death From Verbal Autopsy With Deep Learning Interpretable Methods. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1315–1325.

Blanco, A., Perez-de-Viñaspre, O., Pérez, A., and Casillas, A. (2020d). Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computing Methods and Programs in Biomedicine*, 188:105264.

Blanco, A., Remmer, S., Pérez, A., Dalianis, H., and Casillas, A. (2021b). On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages with Fewer Resources than English. In Angelova, G., Kunilovskaya, M., Mitkov, R., and Nikolova-Koleva, I., editors, *Proceedings of the International Conference on Recent Advances in Natural Language*

Processing (RANLP 2021), Held Online, 1-3 September, 2021, pages 165–172. INCOMA Ltd.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Cheng, Y., Qian, K., Wang, Y., and Zhao, D. (2020). Missing multi-label learning with non-equilibrium based on classification margin. *Appl. Soft Comput.*, 86.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR*, abs/1412.3555.

Dalianis, H. (2018). *Clinical Text Mining - Secondary Use of Electronic Patient Records*. Springer.

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

Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., and Grune, B. (2019). Overview of the CLEF eHealth 2019 Multilingual Information Extraction. *CEUR-WS*.

Dupor, B. and Guerrero, R. (2021). The Aggregate and Local Economic Effects of Government Financed Health Care. *Economic Inquiry*, 59(2):662–670.

Díaz de Ilarraza Sánchez, A., Gojenola Galletebeitia, K., Martínez Unanue, R., Fresno Fernández, V., Turmo Borrás, J., and Padró Cirera, L. (2017). PROcesamiento Semántico textual Avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes MEDicos (PROSA-MED).

European Commission (2020). Horizon 2020 research and innovation funding programme. https://ec.europa.eu/programmes/horizon2020/en/home.

Gallo, F., Seniori Costantini, A., Puglisi, M. T., and Barton, N. (2021). Biomedical and health research: an analysis of country participation and research fields in the EU's Horizon 2020. *European journal of epidemiology*, 36(12):1209–1217.

Gangavarapu, T., Jayasimha, A., Krishnan, G. S., and Kamath, S. (2020). Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems*, 190:105321.

Gobierno de España (2020). Plan de Impulso de las Tecnologías del Lenguaje (Plan TL). https://www.plantl.gob.es/tecnologias-lenguaje/actividades/infraestructuras/Paginas/infraestructuras-linguisticas.aspx.

Goenaga, I., Lahuerta, X., Atutxa, A., and Gojenola, K. (2021). A section identification tool: Towards HL7 CDA/CCR standardization in Spanish discharge summaries. *Journal of Biomedical Informatics*, 121:103875.

Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez, G. G., Viviani, M., and Xu, C. (2020). Overview of the clef ehealth evaluation lab 2020. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 255–271. Springer.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *CoRR*, abs/2007.15779.

Huang, J., Osorio, C., and Sy, L. W. (2019). An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput. Methods Programs Biomed.*, 177:141–153.

Ji, S., Hölttä, M., and Marttinen, P. (2021). Does the Magic of BERT Apply to Medical Code Assignment? A Quantitative Study. *CoRR*, abs/2103.06511.

Lee, C., Fang, W., Yeh, C., and Wang, Y. F. (2018). Multi-Label Zero-Shot Learning With Structured Knowledge Graphs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1576–1585. Computer Vision Foundation / IEEE Computer Society.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li, M., Fei, Z., Zeng, M., Wu, F.-X., Li, Y., Pan, Y., and Wang, J. (2018). Automated ICD-9 coding via a deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4):1193–1202.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with Words at the National Library of Sweden - Making a Swedish BERT. *CoRR*, abs/2007.01658.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., and Al-Garadi, M. A. (2017). Automatic icd-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PloS one*, 12(2):e0170242.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable Prediction of Medical Codes from Clinical Text. In Walker, M. A., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1101–1111. Association for Computational Linguistics.

Névéol, A., Anderson, R. N., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Rondet, C., and Zweigenbaum, P. (2017). CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of

death certificates in english and french. In *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, page 17.

Névéol, A., Cohen, K. B., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., et al. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CEUR workshop proceedings*, volume 1609, page 28. NIH Public Access.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., and Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: opportunities and challenges. *J. Biomed. Semant.*, 9(1):12:1–12:13.

Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., and Zweigenbaum, P. (2018). Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*.

Pérez, J., Pérez, A., Casillas, A., and Gojenola, K. (2018). Cardiology record multi-label classification using latent Dirichlet allocation. *Computer Methods and Programs in Biomedicine.*, 164:111–119.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Qayyum, A., Qadir, J., Bilal, M., and Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Rios, A. and Kavuluru, R. (2018). Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access.

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., and Detmer, D. E. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, 14(1):1–9.

Sammani, A., Bagheri, A., van der Heijden, P. G., Te Riele, A. S., Baas, A. F., Oosters, C., Oberski, D., and Asselbergs, F. W. (2021). Automatic multilabel detection of icd10 codes in dutch cardiology discharge letters using neural networks. *NPJ digital medicine*, 4(1):1–10.

Sankoh, O. and Byass, P. (2014). Cause-specific mortality at INDEPTH health and demographic surveillance system sites in Africa and Asia: concluding synthesis. *Global health action*, 7(1):25590.

Santiso, S., Pérez, A., Casillas, A., and Oronoz, M. (2020). Neural negated entity recognition in Spanish electronic health records. *Journal of biomedical informatics*, 105:103419.

Savelieva, A., Au-Yeung, B., and Ramani, V. (2020). Abstractive summarization of spoken and written instructions with BERT. In Fabbrizio, G. D., Kallumadi, S., Porwal, U., and Taula, T., editors, *Proceedings of the KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption co-located with the 26TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2020), Virtual Workshop, August 24, 2020*, volume 2666 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852. IEEE Computer Society.

Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., et al. (2018). Overview of the CLEF eHealth Evaluation Lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 286–301. Springer.

Tang, D., Qin, B., Feng, X., and Liu, T. (2015). Target-Dependent Sentiment Classification with Long Short Term Memory. *CoRR*, abs/1512.01100.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. (2021). Long Range Arena : A Benchmark for Efficient Transformers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Trigueros, O., Blanco, A., Lebeña, N., Casillas, A., and Pérez, A. (2022). Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. *International Journal of Medical Informatics*, 157:104615.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009). Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Vicepresidencia Tercera, Gobierno de España (2020). Estrategia Nacional de Inteligencia Artificial (ENIA).

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Health Organization et al. (1975). International Classification of Diseases (ICD-10) World Health Organization. *International Classification of Disease and Causes of Death. 9th Revision. Geneva: WHO*.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. (2018). Mining electronic health records (EHRs) A survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40.

Yang, Q. and Shang, L. (2019). Multi-task Learning with Bidirectional Language Models for Text Classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., and Lyman, K. (2017). Learning to diagnose from scratch by exploiting dependencies among labels. *CoRR*, abs/1710.10501.

Yeh, C., Wu, W., Ko, W., and Wang, Y. F. (2017). Learning Deep Latent Space for Multi-Label Classification. In Singh, S. and Markovitch, S., editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2838–2844. AAAI Press.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big Bird: Transformers for Longer Sequences.

Zhang, Y. and Schneider, J. G. (2011). Multi-Label Output Codes using Canonical Correlation Analysis. In Gordon, G. J., Dunson, D. B., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 873–882. JMLR.org.

Zhang, Z., Liu, J., and Razavian, N. (2020). BERT-XML: large scale automated ICD coding using BERT pretraining. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 24–34. Association for Computational Linguistics.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3485–3495, Osaka, Japan. The COLING 2016 Organizing Committee.

Zikos, D. and DeLellis, N. (2018). CDSS-RM: a clinical decision support system reference model. *BMC Medical Research Methodology*, 18(1):137.

A

Appendix

The present Appendix contains the compendium of publications in the suggested reading order. See Section 1.2 for an overview and brief introduction to each work.

Expert Systems With Applications 138 (2019) 112835

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multi-label clinical document classification: Impact of label-density

Alberto Blanco[a], Arantza Casillas[b,*], Alicia Pérez[c], Arantza Diaz de Ilarraza[d]

[a] IXA Taldea. UPV-EHU. Manuel Lardizabal Ibilbidea, 1, Donostia 20018 Spain
[b] Departamento de Electricidad y Electrónica. IXA Taldea. UPV-EHU Barrio Sarrin S/N. Leioa 48940 Spain
[c] Departamento del Lenguajes y Sistemas Informáticos. IXA Taldea. UPV-EHU Rafael Moreno Pitxitxi, 3, Bilbao, 48013, Spain
[d] Departamento del Lenguajes y Sistemas Informáticos. IXA Taldea. UPV-EHU Manuel Lardizabal Ibilbidea, 1, Donostia 20018 Spain

ARTICLE INFO

Article history:
Received 26 March 2019
Revised 4 July 2019
Accepted 21 July 2019
Available online 22 July 2019

Keywords:
Multi-label classification
Document classification
Electronic health records
ICD-10 classification

ABSTRACT

Objective: The goal of this work is the classification of Electronic Health Records using Natural Language Techniques. Electronic Health Records (EHRs) convey valuable clinical information, as diagnoses and patient conditions. We explore several Deep Learning classification models for assigning multiple ICD codes to clinical documents. Within the framework of data mining, the aim of multi-label classification is to associate each instance with a set of labels.

Methods: The multi-label classification is typically carried out based on multiple independent classifiers, in the so-called binary relevance learning approach. Nevertheless, diseases tend to be co-related, independent classifiers are unable to model relationships and do not guarantee the consistency of the predicted label-set. To tackle this, we investigate three Neural Network architectures. We study models that are capable of capturing and modeling label dependencies on the output layer. Moreover, learning from data with low label-density is an inherent challenge in multi-label classification. Thorough experiments were conducted to assess each architecture under different scenarios, varying the language, amount of data and label-density.

Results: The results showed that the Bi-GRU model outperform the DNN and both overcome the baseline (BLR). We observed better results with MIMIC than with Osakidetza corpus. Experimental results showed that as the label-density decreases the prediction task becomes harder. It seems that label-density is very much related to the learning ability of the neural networks and another important factor that affects the inference is the amount of training data.

Conclusions: The contributions of this work are: (a) a comparison among three classification approaches based on Neural Networks on data sets in English and Spanish to cope with the multi-label classification problem and (b) the study of the impact of label-density in prediction capabilities in the multi-label context.

1. Introduction

In the last few years, data is taking a leading role due to the massive generation of documents and information in electronic form. Processing automatically electronic documents offer a reasonable way to help humans in the assimilation of the data. Natural Language Processing (NLP) techniques for Machine Translation (Pathak, Pakray, & Bentham, 2018) or classification (Dwivedi, 2016; Salles, GonÃ§alves, Rodrigues, & Rocha, 2018), for example, trained with large amounts of data to address the problem. This paper faces the challenge of classifying a particular type of documents, Electronic Health Records (EHRs), that is, patient records created on a large scale in the daily routine of the hospitals. Clinical text mining, a sub-field of NLP, is applied to discern knowledge from clinical texts (Dwivedi, 2016; Narducci, Lops, & Semeraro, 2017). Notably, we deal with the classification of EHRs according to the International Classification of Diseases (ICD).

The ICD (Organization, 2004) offers a standard way to encode diseases and other health problems. The currently applied version of the ICD, the 10th version (ICD-10), is arranged, hierarchically, in chapters that distinguish disease types and injuries. The ICD codes consist of a sequence of alpha-numeric characters leading to nearly 70,000 unique ICD codes. So far the ICD is translated into 43 languages and is used to exchange information worldwide.

* Corresponding author.
E-mail addresses: ablanco061@ikasle.ehu.eus (A. Blanco), arantza.casillas@ehu.eus (A. Casillas), alicia.perez@ehu.eus (A. Pérez), a.diazdeilarraza@ehu.eus (A. Diaz de Ilarraza).

https://doi.org/10.1016/j.eswa.2019.112835
0957-4174/

Typically, EHRs are classified with the ICDs and, as a result, comprising a set of classes that characterize, comprehensively, the diseases found. While there is no doubt that ICD codes convey relevant and valuable information, encoding EHRs is a time-consuming task that is carried out by expert clinicians trained on ICD encoding (Dalianis, 2018). Our work aims to explore clinical text mining alternatives to support experts assigning ICD codes to each EHR.

From the text mining point of view, this task consists of classifying a text (the EHR) with a subset of classes from the set $\mathcal{C}$ comprising the standard ICD label-set. Formally, the task is expressed through (1) with X being the representation of the EHR and $\mathcal{C}$ the set of available classes (i.e., the set of codes within the ICD) and $B = \{0, 1\}$ binary information.

$$h : X \rightarrow B^{|\mathcal{C}|} \qquad x \rightarrow h(x) = (b_1, b_2, \ldots, b_{|\mathcal{C}|}) \tag{1}$$

Let us explore the input (x) and the output ($h(x) = y$) of this task. Regarding the input, classical text mining approaches laid on manual hand-crafted symbolic features to represent documents, such as words, part of speech tags, n-grams etc. Nevertheless, these representations tend to suffer from the sparseness of data, out of vocabulary words and generalization issues (Goldberg, 2017). With the development of deep neural networks for NLP (Levy & Goldberg, 2014; Mikolov, Chen, Corrado, & Dean, 2013), the trend is to represent documents as dense vectors ($X = \mathbb{R}^d$). Regarding the output of a multi-label classifier, it is a sub-set of arbitrary size of the label-set and can be represented as $h(\cdot)$ in (1). Note that $B^{|\mathcal{C}|}$ serves to represent arrays of dimension $|\mathcal{C}|$, with the content $b_i \in \{0, 1\}$ interpreted as the membership of the EHR x to the ith ICD-code. That is, $b_i = 1$ if and only if the document x has assigned the class c_i. This way, multi-label classification distinguishes relevant ICD-codes to an input document from irrelevant ones (e.g. $b_i = 0$ would mean that $c_i \in \mathcal{C}$ is not related to the input document x). As a result, $h(x)$ defines a bi-partition of $\mathcal{C}$ for a given document x. The main difference between mono-label and multi-label classification rests on the fact that mono-label classification implies that the classes are mutually exclusive. Accordingly, the size of the label-set is pre-defined to $|\mathcal{C}| = 1$. By contrast, for multi-label classification, a challenge rests on the fact that the size of the label-set related to a given document is variable and unknown in advance. There are widespread applications on text mining aiming at a single label prediction (Kalchbrenner, Grefenstette, & Blunsom, 2014; Kim, 2014; Le & Mikolov, 2014); nevertheless, the EHR classification tackled in this work is beyond that scope.

Often, a multi-label classification task is decomposed as multiple binary classification tasks in the so-called binary relevance approach (Read, Pfahringer, Holmes, & Frank, 2011; Tsoumakas, Katakis, & Vlahavas, 2009). For each document, $|\mathcal{C}|$ binary classifiers are run independently. We can interpret the output by the kth classifier (b_k) as the response to the binary question: "is the kth ICD code c_k relevant to the input EHR?" (with $1 \leq k \leq |\mathcal{C}|$). After this process, the EHR is labeled with the set of classes on which binary classifiers guessed 'yes'. In brief, a binary relevance approach assumes that all labels are statistically independent, hence, disregards label consistency. However, EHRs convey latent constraints, typically different diseases affect different age-ranks (e.g., the subset of relevant diseases to neonatal and pediatric or adult segment is not the same), occupations, social contexts and so on. Label-set consistency seems an aspect to bear in mind when dealing with EHR classification.

Multi-label classification of EHRs with respect to the ICD conveys inherent challenges:

- Supervised classification approaches require a significant sample of pre-classified instances (i.e., EHRs) to learn to predict each class (within the ICD). Nevertheless, one of the core issues of clinical text mining is the lack of corpora available. Besides, not all the diseases have the same prevalence on the population, and hence, there might be classes that are underrepresented leading to data imbalance, a well-known problem (Chawla, Japkowicz, & Kotcz, 2004) to which inference algorithms are sensitive.
- The input is raw text from lengthy documents expressed in free natural language with around a thousand words on average. We should conveniently represent the input EHRs as meaningful dense vectors, x. In this task, we are dealing with plain unstructured, de-identified and disaggregated text (Cohen & Demner-Fushman, 2014; Dalianis, 2018; Williams, Kontopantelis, Buchan, & Peek, 2017). We do not count on other valuable cues (e.g., age or gender fields) to reinforce the representation.

In brief, the goal of this work is to build a robust multi-label EHR classification system adapted to low-density contexts that arise in ICD encoding tasks. To this end, we propose an architecture, based on deep neural networks, an alternative to binary-relevance approach.

Our primary goal is to focus on Spanish, a widely used language for which clinical text mining counts on minimal resources. Despite, we assess the methodology on a noted dataset in English, namely, MIMIC (Johnson et al., 2016; Perotte et al., 2014) in conjunction with a set of EHRs from the Basque Health System, Osakidetza.

2. Background

Our aim focuses on classifying full patient documents EHRs (of hundreds of words) with multiple ICD codes (nearly ten). There are related works that focus on a simplified goal, that is, classifying a diagnostic term (of approximately four words) into an ICD code (Dermouche et al., 2016; Farkas & Szarvas, 2008; Névéol et al., 2017; 2016). While one could figure out the task as a straightforward lookup of the diagnostic term in the ICD dictionary, this was proven ineffective due to the big lexical gap between standard terminology and the terminology employed by doctors in EHRs (Pérez, Atutxa, Casillas, Gojenola, & Sellart, 2018a). Both Cohen and Demner-Fushman (2014) and Dalianis (2018) corroborated lexical complexity underlying EHRs, significant amounts of non-standard abbreviations, common terminology and typos are frequent. Diagnostic term encoding received the attention of CLEF 2016 (Dermouche et al., 2016) and 2017 (Névéol et al., 2017) challenges. A step ahead is made in CLEF 2018 eHealth challenge (Névéol et al., 2018), in which several diagnostic terms (not a single one), are given as input, yet, far from full EHR classification. It is important to note the enormous difference between diagnostic term classification and document classification even though both are classified according to the ICD. For example, Dermouche et al. (2016) dealt with strings of 3.6 words on average and tried to provide an ICD code out of just $|\mathcal{C}| = 60$ classes. In our work, instead, the documents have, hundreds of words, besides, we are dealing with a range of different classes, nearly $|\mathcal{C}| = 1,000$.

We have mentioned the task of one-to-one diagnostic-term encoding, but we need to make a step ahead and delve into the methodology to cope with the multi-label encoding of entire documents. One-to-many or multi-label classification tasks have their nuances regardless of the application (e.g., image or text classification). Reported challenges comprise insufficient and low-density training data (Berndorfer & Henriksson, 2017; Nigam, 2016; Wei et al., 2014; Yao et al., 2017). An open research question in this field is the modeling of class dependencies to predict consistent sets of labels. In this line, Lee, Fang, Yeh, and Wang (2017) paid attention to inter-dependencies between label-sets that had been seen within the training instances and those

that had not. Zhang and Zhou (2006) proposed a multi-label neural network algorithm derived from the popular Backpropagation algorithm, redefining the error function to cope with multi-label learning. Wang et al. (2016) combined RNNs with CNNs with the aim of exploiting latent label dependencies in image classification. The paradigm infers a joint image-label embedding in an attempt to characterize the semantic label dependency as well as the image-label relevance. Baumel, Nassour-Kassis, Cohen, Elhadad, and Elhadad (2018) present a hierarchical RNN based approach to tag documents by identifying the relevant sentences for each label. The hierarchical approach requires an extra preprocessing step, the segmentation of the input texts since the Recurrent Network is not applied over the entire documents. Yeh, Wu, Ko, and Wang (2017) proposed Canonical Correlated AutoEncoder in order to relate features and label domain. They integrated the DNN architectures of canonical correlation analysis and auto-encoder. From the aforementioned works, we observed that the multi-label classification of complete documents requires robust document characterizations to convey all the relevant information that would lead to consistent multi-label generation.

For text multi-label classification within the clinical domain, there are few available corpora to make comparisons. A reference corpus in English is MIMIC III (Johnson et al., 2016; Perotte et al., 2014). Still, our primary interest rests on making progress in processing EHRs in Spanish. To that end, we compiled the so-called Osakidetza corpus, a set of 1610 hospital admissions. Previous authors mined MIMIC in different ways. Li et al. (2017) restricted the dataset to a subset of 13,152 clinical notes comprising the top 10 most frequent ICD codes. Berndorfer and Henriksson (2017) made use of a simplified version of the MIMIC, discarding the labels with a frequency below 50 times leading to a set of 59,531 records with 1301 distinct codes. (Nigam, 2016) focused on two controlled subsets of MIMIC data restricted to, respectively, 10 and 100 ICD codes comprising 31,865 and 44,591 patient records in turn. Logistic Regression was used as the baseline, and they proposed RNN with GRUs (Cho et al., 2014). They concluded that, while with the set of 10 labels the RNN with GRUs outperformed the others, with the set of 100 labels the neural network with GRUs performed worse than the feed-forward network and the standard RNN model. Altogether, we found that MIMIC corpus was often simplified in several ways to make machine learning feasible. The main limitation rests on the reproducibility when it comes to obtaining those sets.

From the related works we learned that in an attempt to convey substantial information from the records, RNN seemed appropriate, but not in all cases, thus, simpler feed-forward networks should not be discarded in the methodology. We observed two gaps in the antecedents: First, the use of binary relevance approaches to tackle ICD multi-label classification disregards the consistency of the label-set. Second, it is common practice to focus just on a subset of highly prevalent labels; thus, low-density cases are not considered. In this work, we delve into both of them.

3. Methods

3.1. Data

Our primary aim is to tackle the multi-label classification of EHRs in Spanish with respect to ICD-10. To that end, a set of EHRs from the Basque public hospital system, Osakidetza, was used. In an attempt to compare our methods to a closely related task of reference, we also employed MIMIC. The main characteristics of these corpora are described in Table 1.

Let us describe, first, the data-sets from the output multi-label perspective. The cardinality of a set $\mathcal{S} = \{(x_j, \mathcal{Y}_j)\}_{j=1}^{N} \subseteq X \times \mathcal{P}(\mathcal{C})$, with $\mathcal{P}(\mathcal{C})$ denoting the power-set of $\mathcal{C}$, is the average number of relevant labels ($|\mathcal{Y}_j|$) associated with the instances (x_j). Formally, this is given in (2), with $b_{i(j)}$ representing the ith membership bit of the instance x_j as in expression (1). $b_{i(j)}$ is equal to 1 if the document x_j has associated the ith ICD-code and assuming that L is the size of the available set of labels (i.e. $|\mathcal{C}| = L$). The cardinality does not take into account the relative number of labels with respect to the entire label-set, by contrast, the density does, as expressed in (3). The behavior of the multi-label learning algorithms is sensitive to the density (Tsoumakas et al., 2009).

Table 1
Data analysis for each data-set $\mathcal{S} = \{(x_j, \mathcal{Y}_j)\}_{j=1}^{N} \subseteq X \times \mathcal{P}(\mathcal{C})$. focusing on both the input and output domains.

		MIMIC	Osakidetza
$\mathcal{S}$	Number of samples (N)	55,172	1610
x_j	Language	English	Spanish
	Total words	73.8×10^6	1.4×10^6
	Distinct words (vocabulary)	225,058	24,116
	Words/Doc (Mean ± St.Dev)	1337 ± 695	866 ± 343
$\mathcal{Y}_j$	ICD version	ICD-9	ICD-10
	Number of labels (L)	6918	974
	Cardinality	11.64	6.95
	Density	0.002	0.007

$$Cardinality(\mathcal{S}) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{L} b_{i(j)} \tag{2}$$

$$Density(\mathcal{S}) = \frac{1}{L} Cardinality(\mathcal{S}) \tag{3}$$

Next, we shall describe the data-sets from the input perspective. Given that for both Osakidetza and MIMIC the information is extracted from raw documents, the size and variability of the documents is a crucial issue. The documents were unstructured, meaning that the doctors did not have fields (such as “Antecedents” or “Analytics”) to arrange their discourse. Simple data cleaning procedures were applied, that is, tokenization, lower-casing, and removal of non-alphanumeric characters. As discussed in Section 2, comparisons with MIMIC are not straightforward since different authors provide results on self-created sub-sets that often are hardly reproducible due to misleading descriptions. For the sake of reproducibility, with this manuscript, we made available the scripts used to compile MIMIC corpus as well as the rest of the software implemented in this work[1]

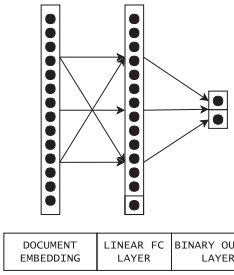
3.2. Classifiers

From the related works, we found interest in exploring three approaches, all neural-based, in increasing complexity order: a Binary Logistic Regresion (BLR); a Deep Neural Network (DNN) and a Bidirectional Gated Recurrent Unit (Bi-GRU). The BLR is an inherently linear classifier that introduces a non-linearity in the form of the Softmax function to compute the loss function. Then, both the DNN and the Bi-GRU include the non-linearities to model the latent representation of the input, being the Bi-GRU, the most complex model. These three approaches are summarized in Fig. 1 and explained, respectively, in Sections 3.2.1–3.2.3.

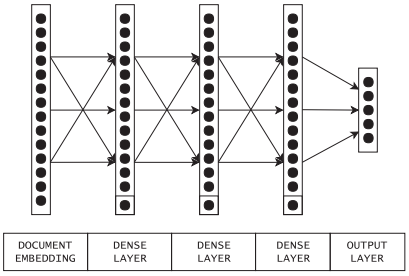
3.2.1. Binary logistic regression

Binary Logistic Regression (BLR) is a parametric classifier that takes a vector of features x describing the input instance and provides, as an output, the prediction through $y = f(X) = W \cdot X + b$. In this case, and from now onwards, the membership referred to

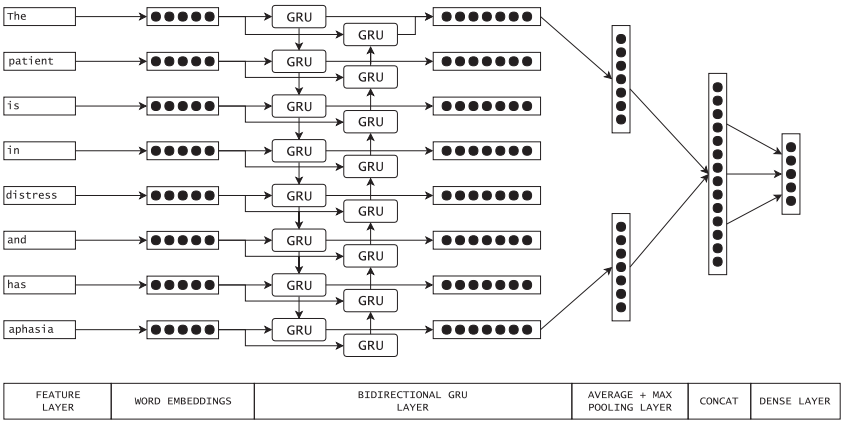
[1] The software implemented in this work is accessible through: http://ixa2.si.ehu.eus/multilabel-consistency-download. With the username `multilabel` and password `consistency`. All parties using this are requested to cite this article.



(a) Binary Logistic Regression architecture used for each class.



(b) Deep Neural Network



(c) Bidirectional Gated Recurrent Unit.

Fig. 1. Architectures used to deal with multi-label classification.

in expression (1) is not directly a bipartition but a real number interpreted as a likelihood with respect to a given class. A threshold function enables to convert the outputs into binary membership to be interpreted as a relevant or irrelevant label. Within the framework of BLR, $f(\cdot)$ is a linear function parametrized through W and b. Training the classifier consists of fitting the function $f(\cdot)$ to the training data adjusting W and b to approach $f(\cdot)$ together with the threshold to the real label membership.

Multi-label classification tasks have been often tackled employing an architecture that entails BLR guessers in parallel (Joachims, 1998; Read et al., 2011; Tsoumakas et al., 2009; Tsoumakas & Vlahavas, 2007). With this approach, a binary classifier is trained for each label. Each classifier is specialized in setting the membership of a particular class (focusing just on one b_i). A limitation of this approach is that a pre-fixed threshold is selected for all classes, typically, $\theta_0 = 0.5 \quad 1 \leq i \leq |\mathcal{C}|$.

Given that this simple approach is the core of neural networks that can cope with non-linear functions, we found this approach a natural baseline for the following approaches. For further details, the reader could turn to Zhang, Li, Liu, and Geng (2018).

Regarding practical aspects, we implemented the BLR employing Tensorflow (Abadi et al., 2015). The network consists of the input layer, a hidden layer, and the output layer, as depicted in Fig. 1(a). Note that for all input EHRs (x_j) a fixed dimensional embedding is used in the feature layer. The output of the classifier in charge of ith ICD code is binary ($b_i \in \{0, 1\}$).

3.2.2. Deep neural network

BLR poses several limitations. On the one hand, the independence assumption (label-sets are generated disregarding consistency). On the other hand, BLR rests on the linearity assumption. Both of them can be addressed by Deep Neural Networks (DNNs) (Goodfellow, Bengio, Courville, and Bengio, 2016, Chapter 6). Thus, this second approach emerges, naturally, as a relaxation of constraints from the baseline.

We implemented a DNN with 3 fully connected hidden-layers as a result the input of the $i-th$ hidden layer is the output of the precedent one $X^{(i-1)}$ again, carrying out a linear transformation, $X^{(i)} = W^{(i)}X^{(i-1)}$ with $X^{(0)}$ denoting the input X. The architecture is depicted in Fig. 1b.

The difference between this DNN and the baseline rests on the output vector. The output of this network comprises as many outputs as distinct classes $|\mathcal{C}| = L$. $Y = [y_{c_1}, y_{c_2}, \ldots, y_{c_L}]$. Nevertheless, a soft-max layer is applied to enable interpreting the result in terms of probabilities. Again, y_{c_i} represents the probability of associating the input document (X) to the ICD code c_i. As with the BLR, the class c_i is assigned to a given document provided the weight exceeds a threshold ($y_{c_i} > \theta_0 \Leftrightarrow b_i = 1$).

3.2.3. Bidirectional gated recurrent unit

The main drawback of the aforementioned DNN lies in the way in which the documents are characterized. Both BLR and DNN consider a document as a vector computed from the continuous bag-of-words (CBOW) representation (Mikolov et al., 2013) of the words that compose the document, disregarding, thus, the sequential structure inherent to the documents. Alternatively, Recurrent Neural Networks offer a characterization of documents that comprise the computation of sequential information and, thus, a document is no longer a function of isolated words, instead, ordered sequences of words are involved (Miftakhutdinov & Tutubalina, 2017; Nam, Kim, Mencía, Gurevych, & Frnkranz, 2014; Nigam, 2016). To cope with text sequentiality we opted for implementing a Bidirectional Gated Recurrent Unit (Bi-GRU) (Cho et al., 2014; Chung, Gulcehre, Cho, & Bengio, 2014).

Like long short-term memory RNNs, GRUs are based on gating mechanisms; nevertheless GRU's memorization mechanism is not decoupled (Chung et al., 2014; Jozefowicz, Zaremba, & Sutskever, 2015) by contrast to other approaches such as LSTM networks (Hochreiter & Schmidhuber, 1997). The previous state is accessed through one gate (r) which is responsible for computing estimated updates to it (h). The updated state (s_t) is computed as an interpolation from previous state (s_{t-1}) and the proposed estimated updates (h) with z being the interpolation parameter controlled by a second gate. The GRU unit is defined as in (4) where z is the update gate, r is the reset gate, h is the hidden state and s_t is the output of the unit at time t. W and U are the parameters and the operation $\odot$ defines the Hadamard product or element-wise multiplication.

$$\begin{aligned} z &= \sigma(x_t U^z + s_{t-1} W^z) \\ r &= \sigma(x_t U^r + s_{t-1} W^r) \\ h &= tanh(x_t U^r + (s_{t-1} \odot r) W^h) \\ s_t &= (1 - z) \odot h + z \odot s_{t-1} \end{aligned} \tag{4}$$

Our implementation, depicted in Fig. 1c, comprises the following layers: the feature layer gets the input document as a sequence of words. The embedding layer converts the tokens into dense representations. The bidirectional GRU layer is responsible for focusing on essential features interpolating previous states and the proposed estimated updates. Sequential memorization mechanisms are not separated. The information flows in two directions (from the beginning to the end of the sentence and back to front). In the pooling layer two techniques were applied, average pooling and max pooling. The resulting two vectors are concatenated to serve as the input of the output layer. The output layer consists of a fully connected layer with a Sigmoid activation function. This layer served to determine class membership.

4. Experimental results

In this section the following research questions are addressed:

- Which of the aforementioned methods performs best for multi-label classification?
- How do these approaches behave on different collections of EHRs (with different languages and sizes)?
- How do these approaches degrade as the density of labels decreases?

4.1. Results

The results are shown in two steps: first, we assessed the impact of individual classifiers (Section 4.1.1) and next, we assessed the impact of varying label density on the performance of our best system (Section 4.1.2).

4.1.1. Set-up stage

To begin with, we explored the performance of the three classifiers proposed in Section 3.2. To this end, we tuned the models on a reduced data-set with just the sub-set of labels that appeared in above 15% of the documents motivated by Berndorfer and Henriksson (2017). For the MIMIC corpus, this selection led to a subset of just 6 classes present in 38,320 documents. For the sake of comparisons with the MIMIC corpus, for the Osakidetza corpus, we restricted the number of classes to the most frequent 6 classes, leading to a subset of 1,186 documents. The reduced sub-sets have a label-density of 0.341 and 0.311 for MIMIC and Osakidetza respectively. Each data-set was randomly divided into two disjoint subsets covering 70% and 30% respectively for training the models and testing them.

With this reduced set, the first aim was to select the most suitable representation for the documents, that is, the x_j feature vector for the jth EHR. While the core features of this work are raw

Table 2
Mean weighted average performance with standard deviation across runs in 5-fold cross validation of each classifier (BLR, DNN, Bi-GRU) for the reduced data-sets. The characters determine the corpus, "M" stands for MIMIC and "O" stands for Osakidetza.

	Precision		Recall		F-Score	
	MIMIC	Osa	MIMIC	Osa	MIMIC	Osa
BLR	66.1 ± 0.2	37.0 ± 1.1	49.3 ± 0.3	18.6 ± 0.4	55.5 ± 0.2	18.7 ± 0.5
DNN	67.5 ± 0.3	45.8 ± 0.3	78.3 ± 0.6	87.7 ± 0.7	72.3 ± 0.1	58.8 ± 0.2
Bi-GRU	77.2 ± 0.3	66.2 ± 0.7	82.6 ± 0.5	80.7 ± 0.9	79.7 ± 0.3	72.5 ± 0.2

word-forms, the word-forms were embedded into dense vectors. To this end we employed FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) with cbow trained from in-domain corpus generating vectors of dimension 300. The training set vocabulary size is 18,258 and 146,831 tokens for Osakidetza and MIMIC corpus, respectively, while the vocabulary size for the test set is 12,342 and 94,907 tokens. The vocabulary sets are disjoint, which leads to Out-of-vocabulary (OOV) tokens, specifically, there are 3071 (24,8% of test vocabulary) and 39,363 (41,4% of test vocabulary) OOV tokens in the test set, for each corpus. At prediction time, when an OOV token is encountered, it is mapped to an OOV distinctive token (<OOV>).

The assessment of the three approaches was carried out using traditional bipartition metrics: precision, recall, f-score. Given that we are dealing with multi-label classification, overall performance is assessed by averaging the metrics across labels. To that end, there are several averaging strategies: micro-, macro-, and weighted-average (Nam et al., 2014; Tsoumakas et al., 2009; Van Asch, 2013). Micro-averaging computes the metrics globally, while macro- calculates the metrics for each label, returning their unweighted mean. While macro-average is insensitive to class skew, micro-average is blind to the classes as it averages the results of the metrics across classes. Weighted average measures the metrics for each label and finds the average weighted by class priors. This way, weighted-average accounts for label imbalance while keeping track of individual classes.

The AUC is also an interesting metric to bear in mind since the classes in these data are highly skewed and AUC is insensitive to class imbalance (Fawcett, 2006).

Next, with this setup and following a repeated 5-fold cross-validation evaluation schema, we assessed the performance of the different classifiers for both MIMIC and Osakidetza corpora. The results attained are shown in Table 2.

From this set of experiments, we learned that both DNN and Bi-GRU approaches outperformed the predictive ability of the baseline (BLR) while Bi-GRU beat the DNN, this all with a with a p-value around 10^{-10}.

When it comes to assessing a model, there is an issue that is often take aside the computational cost involved in the training and prediction process. Table 3 shows the number of parameters, i.e connections between layers plus biases in every layer, of the DNN and Bi-GRU. As shown in Table 3, the number parameters is an order of magnitude higher in Bi-GRU; moreover, the involved time is three orders of magnitude higher. That is, we feel that there is a trade-off between performance and computation cost. Although the DNN performed notably below the Bi-GRU, it may be a compelling approach for tasks with hardware or time constraints.

Table 3
Computational training cost in terms of trainable parameters, training time per epoch and prediction time per step in the 5-fold cross validation assessment.

	Trainable parameters	Training time	Prediction time
DNN	301,000	4,5 s/epoch	30 µs/instance
Bi-Gru	4,414,980	2,915 s/epoch	24,000 µs/instance

4.1.2. Impact of label density on multi-label classification

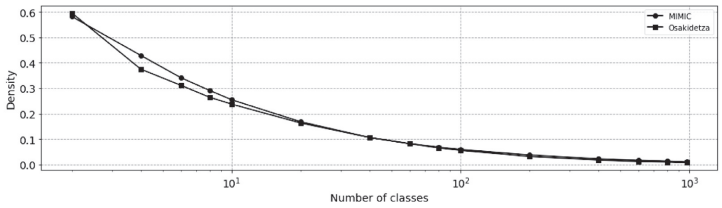
So far we just made use of the reduced sets in an attempt to assess the characterization for the input documents (document-embedding extraction and dimension) together with the choice of classifier (BLR, DNN, Bi-GRU). We concluded that while regarding prediction capacity of Bi-GRU outperforms DNN, it is also true that its computational cost is remarkably higher.

An important point, and the focus of our work, is how the aforementioned techniques deteriorate as the number of classes grows using the entire dataset (presented in Table 1) rather than the reduced one. Needless to say, as we move from the reduced sets towards real sets the complexity of the multi-label classification increases. The issue is not only that the available set of classes ($|\mathcal{C}|$) increases but also that the size of the label-sets tends to increase (and so does the number of different label-sets) and, what is even more critical, the label-density decreases. Besides, the classes are selected regarding decreasing frequency as in Berndorfer and Henriksson (2017). This selection implies that each time we add a new class (ICD-code) into the study to augment the reduced set, the new class has associated fewer instances (EHRs) to learn from the classification pattern. Moreover, this behavior is not linear since we are in a multi-label classification task. All in all, the intuition is that worse prediction ability shall be achieved as the label-density decreases, and so reflects the experimental results attained. First, in Fig. 2a, we showed the relation between the number of classes and label-density. Next, in Fig. 2b and c, we showed the degradation of the performance with both DNN and Bi-GRU, with Osakidetza and MIMIC corpus, respectively. Note that the density is decreasing along the x-axis for the sake of interpretability. Indeed, decreasing density is tightly related to the increase in the number of classes.

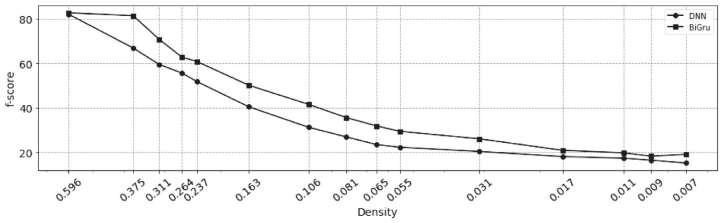
Fig. 2 provided the degradation of f-score varying the label-density. We tested a broad range of L values to gain insights into the impact of label-density. Fig. 2a shows, explicitly, the density decrease rate, with respect to the increase in the number of labels. Note that the relation is similar, ensuring, thus, that the distributions of the labels from both Osakidetza and MIMIC corpus are comparable. The results are shown in Table 4 for two label-densities, corresponding to the size of label-sets $L = 10$ and $L = 100$. These two values of L (with a difference of an order of magnitude) enable the reader compare our results with related works. For example Li et al. (2017) selected a label-set=10; the systems participating in the 2007 Computational Medicine Challenge dealt with 45 classes Pestian et al. (2007), or Pérez, Pérez, Casillas, and Gojenola (2018b) addressed with 124 labels. We discuss the results in the following section.

4.2. Discussion

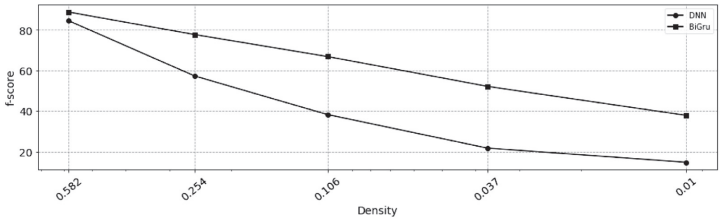
Our work focuses on the multi-label classification of EHRs with respect to ICD. To verify our achievements we experimented with



(a) Label-density varying the number of classes of choice.



(b) Osakidetza corpus.



(c) MIMIC corpus.

Fig. 2. Comparison of the impact of label-set density with MIMIC and Osakidetza corpus for DNN (•) and BiGru (■) models.

both MIMIC and Osakidetza corpora (in English and Spanish respectively). Data analysis showed that label-density of the MIMIC corpus was three times smaller than in Osakidetza. However, there are 30 times as many EHRs in MIMIC as in Osakidetza. Experimental results (Fig. 2) showed the trend of the performance decay as the label-density decreases and we observed that the DNN system attained better results with MIMIC than with Osakidetza. It seems that label-density is very much related to the learning ability of the neural networks and another important factor that affects the inference is the amount of training data.

Table 4
Detailed performance of the best model (Bi-GRU) focusing on two different sizes of label-sets ($L = 10$ and $L = 100$) for two corpora (Osakidetza and MIMIC).

		Osakidetza L		MIMIC L	
Metric	Average	10	100	10	100
Precision	m	55.8	22.3	77.6	56.7
	M	53.9	17.8	76.4	51.6
	w	57.5	29.7	77.6	58.6
Recall	m	68.4	44.9	77.7	62.8
	M	64.2	28.9	76.4	55.3
	w	68.4	44.9	77.7	62.8
F-score	m	61.5	29.7	77.7	59.6
	M	58.0	20.3	76.4	53.0
	w	61.9	34.1	77.6	60.4
AUC	m	75.9	67.9	84.8	79.7
	M	73.4	59.7	84.0	75.9

Concerning the prediction ability, we found that both DNN and Bi-GRU outperformed, notably, the baseline (BLR). Note that the baseline implemented a binary classifier for each class and, hence, the predicted label-set does not take into account latent relations between labels. From Fig. 2 we learned that Bi-GRU outperformed DNN. With this, for a given multi-label set we can figure out the number of instances and label-set density required to achieve particular performance. DNN resulted versatile for both corpora and, overall, achieved good performance.

The approaches found in the antecedents with MIMIC corpus applied different preprocessing steps on the label-set in an attempt to make the task feasible. As MIMIC comes in the form of a relational database, the procedures to extract the input texts have a strong influence and are rarely specified. Besides, the dataset does not yet have predefined training, validation and testing sets, so authors provide results on arbitrary subsets. These procedures are not bright enough and hardly enable conducting fair reproducible results and comparisons.

Again, while the comparisons might result arguably, we would like to provide a review for the sake of completing this study.

Li et al. (2017) restricted the problem to a mono-label (while we are dealing with multi-label) and multi-class classification (with $L = 10$), since they aimed to predict the primary discharge diagnosis. Their approach with CNN achieved 96.11 accuracy and 80.48 weighted F-Score. The difference between mono- and multi-label learning complexity and attained results are worth mentioning.

Berndorfer and Henriksson (2017) employed SVM classifiers to tackle a multi-label problem with L above 1,000, in a binarized fashion with a one-versus-all model per diagnosis code. They combined models trained using both symbolic and dense representations (BoW and word2vec) and achieved 36.95 F-score. Ensemble techniques were used as a means of tackling label-set consistency.

Nigam (2016) explored multi-label problem with $L = 10$ using different approaches based on neural networks and concluded that RNNs with GRUs outperformed the rest attaining 42.03 F-score. In our case, Bi-GRU turned the best approach regarding prediction ability, 77.6 F-score for the case of $L = 10$.

As a secondary contribution of this article, we made available the software used in our work in order to enable reproducible research. Moreover, the classifiers and evaluation modules are also released together with this manuscript.[2]

5. Concluding remarks and future work

This work copes with the multi-label classification of EHRs in English and Spanish with respect to ICD-10. The aim is to assist expert coders in the manual coding process. In this work, we explored natural language processing techniques based on neural network strategies. From the machine learning perspective, this is a difficult task for several reasons. On the one hand, the ICD counts on thousands of classes; hence, we have to face multi-label classification with a big set of classes ($|\mathcal{C}|$ is high). Besides, the number of ICD-codes assigned to each EHR tends to be above 6, but the range of the length of label-sets is diverse (from 1 up to 37). Moreover, the label-set predicted should be consistent, while typical multi-label classification approaches do not bear this issue.

In this work we explored three classification approaches (BLR, DNN, and Bi-GRU) with two corpora, MIMIC dealing with ICD-9 in English and Osakidetza with ICD-10 in Spanish. From this study, we learned that given that multi-label tasks count on scarce instances per class to learn patterns from, data-consistency is an important question to address. Overall, both DNN and Bi-GRU outperformed BLR approach in both corpora. The key issue is that both DNN and Bi-GRU, by contrast to BLR, infer the label-sets jointly and so, address, label consistency. Moreover, Bi-GRU outperformed notably DNN this might be due to the fact that by contrast to DNN, Bi-GRU is able to cope with contextual information in the input. As a secondary contribution, we released the set of programs implemented for this work in an attempt to make further studies on multi-label classification comparable to this one.

For human experts, encoding these lengthy documents entails careful reading EHRs to identify diagnostic terms, procedures, etc. within the full document. These terms are often written in a non-standard language and mapping them into standard terms within the ICD so as to get the corresponding codes is what expert clinicians do in their encoding task. From machine-learning perspective, Bi-GRU mimics, somehow, this process as it focuses on relevant phrases that explain the code-generation process.

The amount of corpora is a significant limitation to gain performance in this task. It seems that inference ability is highly limited by the label-density and number of instances available associated to each label. Nevertheless, our impression is that further research should deepen in label-set consistency rectification approaches as they could leverage the accuracy attained by the inference algorithms. Possibly, the incorporation of the hierarchical structure of the ICD (or alternative ontologies as SNOMED-CT) could help in this line.

Authors contributions

All the authors have contributed equally.

Conflict of interest

The authors declare that there is no conflict of interest.

Ethical

This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgements

The authors would like to thank experts from Galdakao-Usansolo Hospital involved in this research project, notably, Julián Salvador, medical director, and Ana Berta Montero, responsible for clinical documentation.

This work was partially funded by: PROSAMED: TIN2016-77820-C3-1-R 3 and Ikasiker Grant: IkasC_2017_1_0003.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org https://www.tensorflow.org/.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018). Multi-label classification of patient notes: Case study on ICD code assignment. *Workshops at the thirty-second AAAI conference on artificial intelligence*.

Berndorfer, S., & Henriksson, A. (2017). Automated diagnosis coding with combined text representations. *Studies in Health Technology and Informatics, 235*, 201–205.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6*(1), 1–6.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., & Schwenk, H. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics. doi:10.3115/v1/d14-1179.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555.

Cohen, K. B., & Demner-Fushman, D. (2014). *Biomedical natural language processing*. John Benjamins Publishing Company. doi:10.1075/nlp.11.

Dalianis, H. (2018). *Clinical text mining: secondary use of electronic patient records*. Springer.

Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., & Taright, N. (2016). Ecstra-inserm@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. In *CLEF (working notes)* (pp. 61–68).

Dwivedi, A. K. (2016). Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Computing and Applications, 29*(12), 1545–1554. doi:10.1007/s00521-016-2701-1.

Farkas, R., & Szarvas, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics, 9*(Suppl 3), S10. doi:10.1186/1471-2105-9-s3-s10.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies, 10*(1), 1–309.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning*: Vol. 1. MIT Press Cambridge.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[2] The software implemented in this work is accessible through: http://ixa2.si.ehu.eus/multilabel-consistency-download. With the username `multilabel` and password `consistency`. All parties using this are requested to cite this article.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Machine learning: ECML-98* (pp. 137–142). Springer Berlin Heidelberg. doi:10.1007/bfb0026683.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, 160035. doi:10.1038/sdata.2016.35.

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International conference on machine learning* (pp. 2342–2350).

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. Association for Computational Linguistics. doi:10.3115/v1/p14-1062.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics. doi:10.3115/v1/d14-1181.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196).

Lee, C. W., Fang, W., Yeh, C. K., & Frank Wang, Y. C. (2018). Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1576–1585). ISO 690.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Li, C. Y., Konomis, D., Neubig, G., Xie, P., Cheng, C., & Xing, E. P. (2017). *Convolutional neural networks for medical diagnosis from admission notes* arXiv:1712.02768.

Miftakhutdinov, Z., & Tutubalina, E. (2017). Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. *CLEF Google Scholar*.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of workshop at ICLR, 2013*.

Nam, J., Kim, J., Mencía, E. L., Gurevych, I., & Frnkranz, J. (2014). Large-scale multi-label text classification — revisiting neural networks. In *Machine learning and knowledge discovery in databases* (pp. 437–452). Springer Berlin Heidelberg. doi:10.1007/978-3-662-44851-9_28.

Narducci, F., Lops, P., & Semeraro, G. (2017). Power to the patients: The healthnet-social network. *Information Systems, 71*, 111–122.

Névéol, A., Anderson, R. N., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., et al. (2017). CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in english and french. In *CLEF 2017 evaluation labs and workshop: Online working notes, ceur-ws* (p. 17).

Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., et al. (2018). Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. *CLEF 2018 evaluation labs and workshop: Online working notes, ceur-ws*.

Nigam, P. (2016). *Applying deep learning to ICD-9 multi-label classification from medical records*.

Névéol, A., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., et al. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CEUR workshop proceedings: Vol. 1609* (pp. 28–42).

Organization, W. H. (2004). *International statistical classification of diseases and related health problems*: Vol. 1. World Health Organization. doi:10.1007/978-0-387-79948-3_3055.

Pathak, A., Pakray, P., & Bentham, J. (2018). English–mizo machine translation using neural and statistical approaches. *Neural Computing and Applications*. doi:10.1007/s00521-018-3601-3.

Pérez, A., Atutxa, A., Casillas, A., Gojenola, K., & Sellart, A. (2018). Inferred joint multigram models for medical term normalization according to ICD. *International Journal of Medical Informatics, 110*, 111–117. doi:10.1016/j.ijmedinf.2017.12.007.

Pérez, J., Pérez, A., Casillas, A., & Gojenola, J. (2018). Cardiology record multi-label classification using latent Dirichlet allocation. *Computer Methods and Programs in Biomedicine, 164*, 111–119.

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2014). Diagnosis code assignment: Models and evaluation metrics. *Journal of the American Medical Informatics Association, 21*(2), 231–237. doi:10.1136/amiajnl-2013-002159.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., & Cohen, K. B. (2007). A shared task involving multi-label classification of clinical free text. In *Proceedings of the workshop on bioNLP 2007: Biological, translational, and clinical language processing*. In *BioNLP '07* (pp. 97–104). Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1572392.1572411.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning, 85*(3), 333–359. doi:10.1007/s10994-011-5256-5.

Salles, T., GonÃ§alves, M., Rodrigues, V., & Rocha, L. (2018). Improving random forests by neighborhood projection for effective text classification. *Information Systems, 77*, 1–21.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In *Data mining and knowledge discovery handbook* (pp. 667–685). Springer US. doi:10.1007/978-0-387-09823-4_34.

Tsoumakas, G., & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *Machine learning: ECML 2007* (pp. 406–417). Springer Berlin Heidelberg. doi:10.1007/978-3-540-74958-5_38.

Van Asch, V. (2013). Macro-and micro-averaged evaluation measures. *Tech. Rep*.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE. doi:10.1109/cvpr.2016.251.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., & et al. (2014). CNN: Single-label to multi-label. arXiv:1406.5726.

Williams, R., Kontopantelis, E., Buchan, I., & Peek, N. (2017). Clinical code set engineering for reusing EHR data for research: A review. *Journal of Biomedical Informatics, 70*, 1–13. doi:10.1016/j.jbi.2017.04.010.

Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., & Lyman, K. (2017). Learning to diagnose from scratch by exploiting dependencies among labels. arXiv:1710.10501.

Yeh, C.-K., Wu, W.-C., Ko, W.-J., & Wang, Y.-C. F. (2017). Learning deep latent space for multi-label classification. In *Conference on artificial intelligence* (pp. 2838–2844).

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., & Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science, 12*(2), 191–202. doi:10.1007/s11704-017-7031-7.

Zhang, M.-L., & Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering, 18*(10), 1338–1351. doi:10.1109/tkde.2006.162.

Computer Methods and Programs in Biomedicine 188 (2020) 105264

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity

Alberto Blanco*, Olatz Perez-de-Viñaspre, Alicia Pérez, Arantza Casillas

IXA Taldea. UPV-EHU, Manuel Lardizabal Ibilbidea, 1, Donostia 20018, Spain

ARTICLE INFO

Article history:
Received 1 August 2019
Revised 26 November 2019
Accepted 5 December 2019

Keywords:
Electronic health record
International classification of diseases
Multi-label classification
Recurrent neural networks
Contextual embeddings
Label-granularity

ABSTRACT

Background and objective: This work deals with clinical text mining, a field of Natural Language Processing applied to biomedical informatics. The aim is to classify Electronic Health Records with respect to the International Classification of Diseases, which is the foundation for the identification of international health statistics, and the standard for reporting diseases and health conditions. Within the framework of data mining, the goal is the multi-label classification, as each health record has assigned multiple International Classification of Diseases codes. We investigate five Deep Learning architectures with a dataset obtained from the Basque Country Health System, and six different perspectives derived from shifts in the input and the output.

Methods: We evaluate a Feed Forward Neural Network as the baseline and several Recurrent models based on the Bidirectional GRU architecture, putting our research focus on the text representation layer and testing three variants, from standard word embeddings to meta word embeddings techniques and contextual embeddings.

Results: The results showed that the recurrent models overcome the non-recurrent model. The meta word embeddings techniques are capable of beating the standard word embeddings, but the contextual embeddings exhibit as the most robust for the downstream task overall. Additionally, the label-granularity alone has an impact on the classification performance.

Conclusions: The contributions of this work are a) a comparison among five classification approaches based on Deep Learning on a Spanish dataset to cope with the multi-label health text classification problem; b) the study of the impact of document length and label-set size and granularity in the multi-label context; and c) the study of measures to mitigate multi-label text classification problems related to label-set size and sparseness.

1. Introduction

Methodical documentation of healthcare data is fundamental for public health. The **International Classification of Diseases (ICD)** is the standard diagnoses coding system for **Electronic Health Records (EHR)** classification. ICD serves, worldwide, for epidemiology, health management and documentation purposes. Over time, several versions have been developed, being the ICD-10th the current version. Regarding the hospital network associated with the Spanish "Ministerio de Sanidad, Servicios Sociales e Igualdad", from January the 1st 2016, the clinical modification of the ICD-10th is the reference version, adopting the Spanish translated CIE-10-ES variant as the coding standard. The ICD-10 is designed as an alphanumeric code and it is arranged hierarchically [1]. Each code is built by a set from 3 to 7 alphanumeric characters as shown in Fig. 1.

In this paper we tackle the **task** of automatically coding the diagnostic terms present in a free-text medical record according to the ICD coding system. The task is framed within the Natural Language Processing (NLP) field. The purpose is to determine which classes are present in the input text. Our approach rests on machine learning, specifically on supervised multi-label classification.

Classification based solely on text is an open **challenge** in artificial intelligence [2–4]. We aim to solve a text classification problem on medical free-text, EHRs that present medical jargon and clinical-specific language. Furthermore, EHRs often contain abbreviations (frequently non-standard), and misspellings are also common. The length of the texts plays an important role, here we face

* Corresponding author.
E-mail address: ablanco061@ikasle.ehu.eus (A. Blanco).

https://doi.org/10.1016/j.cmpb.2019.105264
0169-2607/

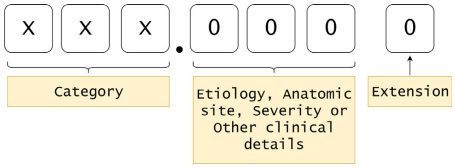


Fig. 1. ICD-10 code structure.

a broad spectrum, ranging from a few words to several tens of lines. EHRs seldom express clinical diagnoses as in the standard ICD.att

An EHR could entail many diagnostics henceforth, multiple ICD labels should be assigned. This task, **Multi-label Classification**, can be seen as a multi-class classification (not binary) task in which the classes are not mutually exclusive. Multi-label classification tends to be, by far, more challenging than mere multi-class classification. Its complexity lies in the exponential growth of label combinations. Note, as well, that the number of labels associated to each EHR is variable.

Multi-label classification can be tackled with the so-called binary relevance approach. This simplistic approach consists of using as many binary classifiers as ICD codes to determine if each ICD code is present or absent from the EHR. The drawback of this approach rests on the fact that the model is not able to capture label-dependencies. While some diagnostics are prone to co-appear others are incompatible. Learning label-dependencies is crucial to this task. To this end, we explore approaches based on Deep Learning [5–7].

The contribution of this work is to explore the impact of dataset characteristics, such as the characterization of the input text focused on (either full document or a part of it), on the predictive ability of the multi-label neural models and also to assess the performance with respect to label-set cardinality and granularity. We deal with real EHRs from Osakidetza (the Basque Public Health System) written in Spanish.[1]

2. Related work

Text classification of EHRs is a demanding task, hence most works have focused on short English texts, though on this work we deal with novel challenges including long EHRs written in Spanish with thousands of words.

Multi-label classification is a challenging task, especially when the number of labels is high [8–10]. The binary relevance approach transforms the multi-label problem in multiple binary classification problems [11], but disregard the dependencies among labels. Several works have addressed the EHRs classification according to the ICD [12–15]. Yet, little attention was paid to dense features and to the approaches that could take advantage of them. Furthermore, much uncertainty still exists about the inter-dependency of labels, that could enhance the prediction performance avoiding incongruities such as, for example, assigning an adult-specific disease simultaneously with a childhood condition. On this work, we tackle the model and capture of label dependencies through Deep Learning models, leveraging the dense output layer with Sigmoid activation function.

The text classification field has leapt forward, from linear and probabilistic models over hand-crafted engineered features [16,17] to non-linear Neural Network models and end-to-end learnt inherent high-level text representations. It is shown good performance with NN [18], as Convolutional Neural Networks [19], Recurrent Neural Networks [20] and Bidirectional Long Short-Term Memory [21].

Methods of **meta-embeddings** aim to conduct a complementary combination of information from an ensemble of distinct word embeddings to yield an embedding set with enhanced quality and characteristics of the semantics captured. Yin and Schütze [22] presented, among others, the "concatenation" method, wherein the meta-embedding is the concatenation of several embeddings. Coates and Bollegala [23] assured that direct averaging of embedding can provide an approximation of the efficiency of concatenation without increasing the dimension of the embeddings.

Context representations are vital to NLP tasks such as text classification. To alleviate this weakness present in generic word embeddings the **contextual embeddings** emerged. Melamud et al. [24] presented an unsupervised model for learning context embedding of wide contexts of sentences using bidirectional LSTMs. These embeddings are dependent on the entire corpus from which they were inferred and carry reinforced contextual meaning. The ELMo [25] and BERT [26] have become state-of-the-art in contextual word representations. Much uncertainty still exists about the advantages of applying meta and contextual embeddings over the standard options for clinical text classification tasks, and we have found that the contextual embeddings may give an extra edge on the ICD classification.

In the automatic ICD coding, there are also works that point towards the Neural Network trend but seems to fall short on the field. These models manage to handle large amounts of text through a dense representation of words. Nigam [27] took advantage of Recurrent Neural Networks to perform multi-label classification. Both works were carried out with discharge summaries from the MIMIC-III [28] corpus. Recently, this task has gained more attention through the CLEF eHealth evaluation labs. Suominen et al. [29] presented an overview of the sixth annual edition. The goal of one of the tasks is to automatically assign ICD-10 codes to few words length texts from free-text descriptions of causes of death as reported by physicians [30,31]. The task is similar to what we have presented on this work with the Diagnostic input perspective, and the finding is that the performance of the classifiers could be improved employing the full documents.

Spanish NLP is under strong growth, among others, driven by the Plan de Tecnologías del Lenguage.[2] EHRs in Spanish are currently being collected [32,33], as well as complimentary corpora including abstracts [34]. These data sets enable to develop several tasks e.g., Negation Extraction [35], Extraction of Adverse Drug Reactions [36], Text Classification [30,31,37], and Negation Cue Detection [38].

3. Methods

We explored four unique RNN model instances plus the baseline model, a Feed Forward Neural Network with Neural-Net Language Model (NNLM) as the text representation layer. The core architecture is a *Bidirectional Recurrent Neural Network with GRU* units and pooling techniques [39] (explained in Section 3.1). The cornerstone of the model is the word embedding layer, as it is responsible for the expressiveness of the input. Thus, we explored three variants: standard embeddings, meta-embeddings and contextual embeddings (explained in depth in Section 3.2). Together

[1] The dataset contains sensitive, confidential data, and therefore can not be released.

[2] https://www.plantl.gob.es/tecnologias-lenguaje/actividades/infraestructuras/Paginas/infraestructuras-linguisticas.aspx.

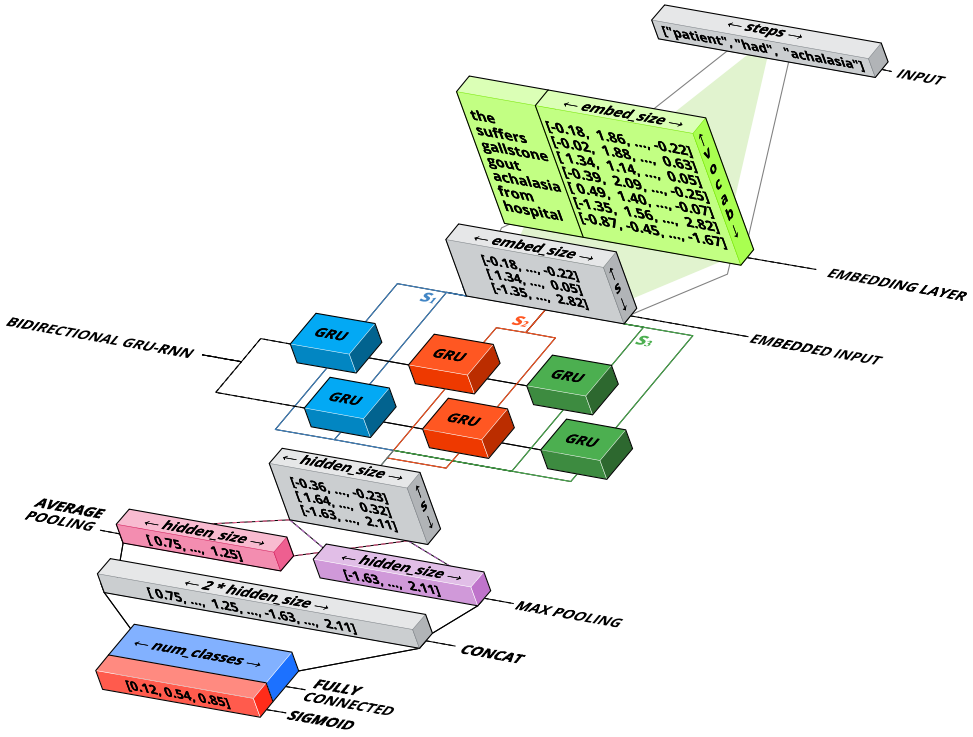


Fig. 2. Architecture: Bidirectional RNN with GRU units and pooling model.

with this work, in an attempt to promote reproducibility, we released the software package that we implemented.[3]

3.1. Bidirectional recurrent neural network with GRU units and pooling

We applied a Bidirectional layer with GRU units, which leverages sequences of text in forward and reverse order with separate hidden states, and whose mathematical formulation for the forward and backward hidden state and its combination is shown in (1).

$$\begin{aligned} \overrightarrow{h^{(t)}} &= \sigma(\overrightarrow{W}x^{(t)} + \overrightarrow{V}\,\overrightarrow{h}^{(t-1)} + \overrightarrow{b}) \\ \overleftarrow{h^{(t)}} &= \sigma(\overleftarrow{W}x^{(t)} + \overleftarrow{V}\,\overleftarrow{h}^{(t-1)} + \overleftarrow{b}) \\ h^{(t)} &= [\overrightarrow{h}^{(t)}, \overleftarrow{h}^{(t)}] \end{aligned} \tag{1}$$

The parameters are the weight matrices $[\overrightarrow{W}, \overleftarrow{W}]$ and $[\overrightarrow{V}, \overleftarrow{V}]$, and the bias terms $[\overrightarrow{b}, \overleftarrow{b}]$. The hidden-states are computed through the non-linear activation (σ) applied to the weighted sum between previous hidden-states $[\overrightarrow{h}^{(t-1)}, \overleftarrow{h}^{(t-1)}]$ and current input $(x^{(t)})$ with their corresponding matrices. Then, both hidden states are combined with concatenation to provide the resulting hidden state (h^t).

The output of the Bidirectional RNN layer could be fed to the dense layer. However, this can be computationally challenging, due to the high number of parameters. Learning a classifier with too many parameters can be unwieldy, and can also be prone to overfitting. A popular technique to deal with the high dimensionality of the Bidirectional RNN layer output is Pooling [40]. We applied average and max-pooling, known as 1-dimensional global pooling. The pooled features are concatenated and fed into a final fully-connected layer. This layer is responsible for computing the probability estimation of the labels i.e. ICD codes. Fig. 2 shows the full architecture of the Bidirectional Recurrent Neural Network with GRU units and pooling techniques, i.e. BiGru. The figure shows a forward pass for an example text. The output of the Sigmoid function is the probability estimation of each label. The depth of every layer indicates the batch size. The Recurrent layer is unrolled, so $s_i \forall i \in s$ brings the embedded representation of the input token $\{s_1 = emb(\text{“patient”}),\ s_2 = emb(\text{“had”}),\ s_3 = emb(\text{“achalasia”})\}$.

The BiGru model can handle all the labels at once, instead of following a binary relevance approach, training independent classifiers for each label. The final dense layer is able to capture and model the label dependencies, producing a non-mutually exclusive

[3] The software is available at http://ixa2.si.ehu.es/prosamed/cmpICD_soft and can be downloaded with user `CMPB` and password `IXAcmpb`. Provided that the software is used anyhow, this article should be cited.

probability estimation for each label with the Sigmoid activation function [41].

3.2. Comprehensive input characterization: embedding layer variations

A comprehensive input characterization is crucial for attaining competitive performance. In the training stage, the embedding layer holds more than 90% of the model's complexity in terms of parameter count. What is more, the predictive capacity rests on the ability of the model to extract knowledge from the source provided in the input stage. Thus, we paid special attention to this layer. The embedding layer from the Fig. 2 shows just a vanilla embedding layer that we enhanced later. Indeed, in this work we explored three variations of the embedding layer: i) Standard embeddings. ii) Meta embeddings (Sections 3.2.1 and 3.2.2). iii) Contextual embeddings (Section 3.2.3)

Moreover, according to Yin et al. [42] and Coates and Bollegala [23], different pre-trained word embeddings have substantial differences in quality and characteristics of the word representations. The consequence is some word embeddings performing better on some tasks than in others. Bearing all this in mind, in addition to a standard pre-trained embedding, we tried *meta-embeddings*, which are ensemble approaches (embedding concatenation and blending) with the hope to get an embedding set with the improved overall quality.

We turned to embeddings derived from fastText [43] as the standard embeddings setup. As for meta-embeddings setup, we employed fastText, Word2Vec [44] and GloVe [45]. Every embedding set is trained on the same corpus, the Spanish Billion Word Corpus [46].

3.2.1. Embedding concatenation

The meta-embedding is computed as the concatenation of word embeddings, based on the work by Yin and Schütze [22]. Before the concatenation, each embedding set must be L2-normalized [6], so that all the values are in the range $[-1, 1]$ and, therefore, every set contributes equally.

The dimensionality of the resulting meta-embeddings is $\hat{d}_{s_k} = d_{s_1} + \ldots + d_{s_i} + d_{s_n}$ with d_{s_i} being the dimension of the ith set concatenated. It is important to note that the model's complexity increases with each added embedding set, as it increases the dimension of the features of the embedding layer.

3.2.2. Embedding blending

The meta-embedding variant is computed as the average of the embeddings involved, based on the work by Coates and Bollegala [23]. Note that even having embedding sets with matching number of dimensions ($d_{s_i} = d_{s_j} \forall\ i, j$), each dimension among embeddings is not related. In any case, averaging can provide an approximation of the performance of concatenation without the expense of increasing the dimension [23].

3.2.3. Contextual embeddings

Recently, approaches that improve the semantic word representation by leveraging the context to encode syntactical meaning and handle polysemy are pushing the state-of-the-art. Regular word embedding techniques use all the occurrences of a word to extract a joint representation. However, depending on the context, words could have different meanings. Recent models exploit this reasoning and propose contextual word embeddings. There is no longer a lookup table between words and dense representations. Instead, the word embedding is computed on the fly, taking advantage of the context.

Embeddings from Language Models (ELMo) [25] representations are obtained from a bidirectional Language Model (biLM) that has recently produced state-of-the-art results in several NLP tasks like Coreference Resolution [47] or Natural Language Inference and Sentiment Analysis [25]. The embedding for a given word varies from one sentence or document to another with its context. As it cannot be pre-computed, the embedding computation is done computing a forward propagation of the model for each token of each input sequence [48].

Table 1
Statistical characterization of the Osakidetza dataset and every perspective. The input can comprise either a small part of the document denoted as "Diagnostic" or the entire "Document" (full EHR). The output (ICD code) can be explored at different granularity levels: Chapter, Block, Full-code.

	Corpus	EHRs		
$\mathcal{S}$	Samples	10,707		
	Input	Diagnostic	Document	
$\mathcal{X}$	Vocabulary size	12,811	60,197	
	Words per doc	37.9 $\pm$ 73.8	770.5 $\pm$ 351.2	
	Output	Chapter	Block	Full
$\mathcal{Y}$	Distinct labels	24	991	3572
	Avg. Cardinality	3.7	5.4	5.5

4. Experimental framework

4.1. Data

The datasets used in our experiments consist of EHRs written in Spanish from the Basque public health system (Osakidetza). Specifically, emergency services discharge summaries from hospitals. The EHRs are not structured and were not written using templates with sections. Table 1 introduces the details of the dataset used. There are 10,707 EHRs. As revealed by the table, we considered several perspectives of the **dataset** by varying two factors, the input and the output explained in what follows.

According to the **input**, the shift consists of the retained proportion of text from the full EHR. Our aim is to determine whether the neural models were able to extract the information from entire documents or, could be benefited from small pieces of text conveying meaningful information. As a result, we explored, on the one hand, the full EHR (referred to as "Document") and on the other hand a short part of the document (referred to as "Diagnostic") as shown in Fig. 3. Note that the mean text length of the Diagnostic perspective is $\approx$ 38 words, while for the Document perspective, it

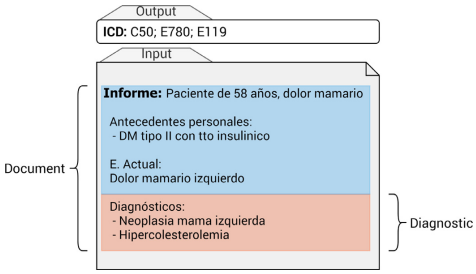


Fig. 3. Health record from the Osakidetza datasets. The ICD codes are listed right to the ''ICD:'' keyword. Every character after the ''Informe:'' keyword is part of the text of the report.

Table 2
Evaluation results for the {***Corpus:*** *Big,* ***Out:*** *Full-code}* varying the input (full document or diagnostic section).

Inp	Model	Characterization		Precision	Recall	F-Score
Document	NNLM	Document embeddings		50.968	30.976	31.513
	BiGru	Standard fastText emb.		51.485	55.825	53.307
		Meta embeddings	Avg	64.824	57.416	59.533
			Conc	65.345	54.428	58.576
		Contextual ELMo emb.		**67.283**	**60.490**	**63.165**
Diagnostic	NNLM	Document embeddings		54.971	34.974	39.257
	BiGru	Standard fastText emb.		51.257	49.785	49.838
		Meta embeddings	Avg	58.482	50.860	53.864
			Conc	**58.752**	50.688	53.992
		Contextual ELMo emb.		56.648	**52.472**	**54.301**

rises to $\approx$ 770, but in both cases, the standard deviation is high, as shown in Table 1.

Regarding the **output**, the shift is in the granularity of the labels, with a three-level alternative taken into account, as follows: The *"Full-code"* level preserves the original code (e.g., "`M1A.1421`": *Chronic gout, lead induced in hand left with tofus*), the *"Block"* level keeps the first three characters (e.g., "`M1A`": *Chronic gout*), and the *"Chapter"* level keeps only the first character (e.g., "`M`": *Diseases of the muskuloskeletal system and connective tissue*).

Fig. 4 shows the resulting label distributions for each output alternative of the Osakidetza datasets. The diseases are not uniformly distributed, as there are frequent and rare conditions. Indeed, the class imbalance is one of the challenges of ICD classification, as machine learning models struggle to handle classification tasks with classes that present large disparities in prevalence. To cope with the high imbalance and high scarcity of some labels Dermouche et al. [49] set a threshold of minimum occurrences per label i.e. keeping only those labels that appear in more than 15 records. Following similar reasoning, and to keep consistency across perspectives, we set a relative threshold, based on the percentage of appearances. Specifically, we keep only those labels that appear in at least 5% of EHRs. The relative threshold enables to keep every perspective label-set coherent concerning the label distribution and minimum class support, while leaving enough samples with each label to evaluate on the test set.

4.2. Results

The experimental set is designed to provide a full range of insights about the application of neural networks to an EHR-ICD based multi-label Spanish text classification task. To that end, we have explored the performance of the 5 models over 6 dataset perspectives.

4.2.1. Assessing the models and the impact of the input text

To evaluate the impact of document length, we make use of the Diagnostic and Document perspectives. Here, the intuition is that extracting the most relevant part of the documents may improve the results by focusing the attention and preventing long-range sequence problems [50,51], but also may harm, due to loss of information, like the mention of symptoms or drugs, on the discarded text.

Table 2 assess the models with either diagnostic or full document as input and full-code labels as output.

Comparing the **models and representation**, we can derive the best performing approach. The baseline is outperformed by every model by a noteworthy amount, besides, the BiGru ELMo outperformed the others in terms of F-score. The BiGru with standard embeddings obtained average results, and the meta-embeddings surpassed the model using just standard one-sided embeddings derived from fastText.

Comparing the amount of information conveyed by the **input text** and relevant to the models to make their predictions, the results are of much interest. All the recurrent models are favoured when providing the full document as input (just the baseline is superior for the diagnostic input). Indeed, the mean difference in terms of F-score in favour of the document input is around 5 points. These results could lead to an extensive discussion. We now bestow an argument: the recurrent models can take advantage of longer sequences with more information and larger vocabulary successfully. Notwithstanding, if the model is not suitable for sequential data (like the baseline model, NNLM, which is not recurrent), those long sequences and large vocabulary weaken the performance, allowing an improvement by keeping shorter text fragments. With these results in mind, we recommend text summarization or similar techniques for extracting the most relevant fragments of texts as a mitigation measure in case of non-recurrent models for document classification tasks. The results attained by the best models for each document length (i.e. Diagnostic and Document perspectives) are marked in bold.

4.2.2. Assessing the impact of label-set size and granularity

Regarding the influence of label-set size and granularity, the intuition is that as the label-set size decreases the performance increases, as the inherent difficulty of the problem diminishes. Besides, it is interesting to explore if the granularity, the degree of label detail, by itself, has an impact on performance.

It is important to remember that due to the relative threshold for label-set reduction, the block label-set of the big dataset has more labels than the full-code label-set, as shown in Fig. 4. Hence, this scenario allows us to check the impact of label-granularity alone.

Due to the high number of entries that a table would require ($n = 30$), and for the sake of clarity, we have chosen to show the outcome of this experiment by a line plot. Fig. 5 shows the F-score (y-axis) for the full-code, block and chapter labels (x-axis) achieved by each of the five models explored for both input perspectives, and we can observe similar behaviours.

Focusing on the document input, we can observe that the behaviour for every model is also similar, improving results as the granularity decreases. One key finding is that the granularity has an impact alone. With less granularity, the performance increases, even with more number of labels. This finding is depicted by the situation between the full labels ($n = 16$) and the block labels ($n = 19$), where with the block labels the performance improves despite having 3 more labels. This suggests that is possible to get models performing better with the same number of labels by just decreasing the label granularity.

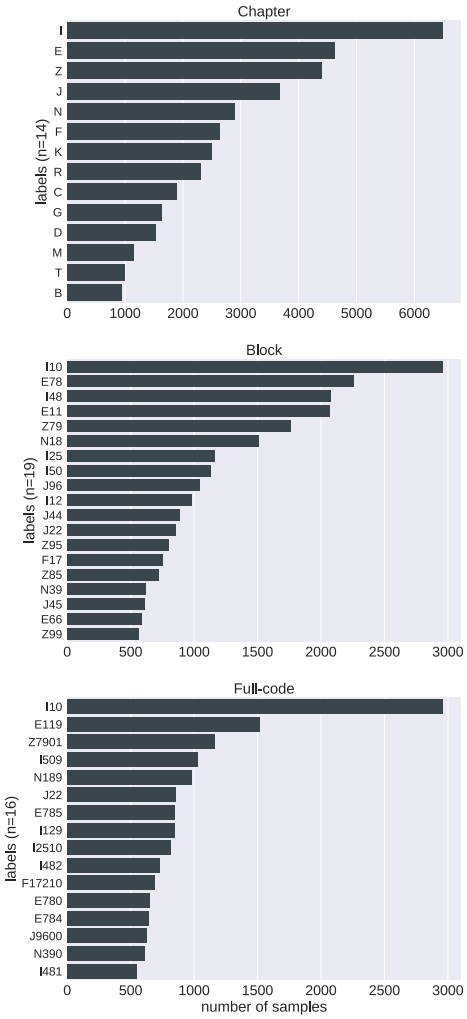


Fig. 4. Label distributions obtained after the reduction of the label-set size with a relative threshold of 5% carried for each output perspective of the Osakidetza datasets.

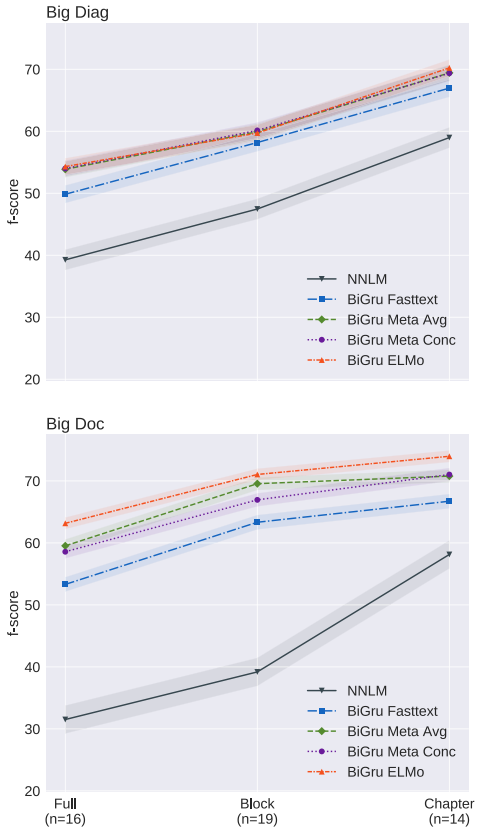


Fig. 5. Line plots for performance comparison among the full-code, block and chapter labels for both perspectives: diagnostic (top), document (bottom).

4.2.3. Discussion

With this work we gained the following **insights**: Despite is a difficult task, Deep Learning recurrent models exhibit strong predictive capabilities and can be enhanced by more robust text representation techniques such as the meta or contextual embeddings.

We argue that our experimental results throw one key **finding**: the granularity of the labels alone has an impact on performance. The significance lies in the possibility of performance improvement by reducing the granularity without reducing the label-set size.

BiGru powered by ELMo is the dominant model in practically every situation from both the input and output perspectives (shown in Table 2 and Fig. 5). Accordingly, a **per-class evaluation** on the best-performing dataset perspective is shown in Fig. 6.

Draw attention to the fact that the worse performing label T (“*Injury, poisoning and certain other consequences of external causes*”) gets 41.7% F-score, while the best-performing label C (“*Neoplasms*”) reaches an outstanding 91%. Half of the labels are above 70% and the $\approx$ 30% of labels are above 80%.

To assess the stability of the models and the statistical significance of the results, we performed five runs repeating the experimental set with random seeds and found that Stdev. among runs remained under 0.5 for precision and recall and 0.25 for F-score for every model and setup, which means that the given experimental results are both reproducible and representative.

Fig. 6. Per-class evaluation of BiGru ELMo based on F-score and class frequency for the {***Inp:*** *Document*, ***Out:*** *Chapter}* subtask.

5. Conclusions

We presented a set of Deep Learning methods to tackle the NLP challenge of multi-label text classification with medical free-text: EHRs written in Spanish with datasets from the Basque Country Health System and classified according to the ICD. Each EHR is assigned multiple ICD codes, leading to multi-label classification of text.

In this work we turned to deep neural models and we found that contextual information conveyed by the BiGru ELMo achieved competitive results. BiGru, by contrast to main approaches seen in the literature, has a mechanism to cope with label co-apparitions and regard diseases as related.

We wondered if the neural models were able to extract the information from entire EHRs of nearly a thousand words or could be boosted by selecting a small though representative section (diagnoses). Experimental results showed that it is worthy providing the model with the full document as it might convey meaningful information. Particularly, BiGru powered with contextual embeddings form ELMo(BiGru+ELMo) outperformed the rest of the models explored. In fact, BiGru+ELMooutperformed every model in all the setups. The difficulty of correctly predicting a label is not the same across labels. A per-class evaluation revealed the competitive performance of this approach on minority classes. That is, BiGru+ELMo resulted robust regarding the class imbalance and, obviously, leveraged frequent ICDs.

Finally, we explored the performance attained varying the output label granularity (fully-specified code, block, chapter) and label-set cardinality (from 14 to 19). This is of interest to decide whether to create a fully automatic ICD classification engine or, depending on the performance required, make the decision to let the model just predict a higher order in the hierarchy.

There are several open directions for future work. First, our models leverage ELMo based contextual embeddings, but there are other novel approaches to contextual embeddings based on Language Models, like BERT [26]. Second, the core architecture of this work is the Recurrent Neural Network, but there are other intriguing architectures like Convolutional Neural Networks, especially Capsule Network [52] or the architecture behind BERT, the new RNN alternative promising approach called Transformer [53]. Third, the methods to address the relation among labels, such as statistical driven approaches (e.g., correlation analysis [54]) and strategies leveraging the hierarchically structured ICD and related ontologies (e.g., Hierarchical Multi-label Classification [55] and SNOMED-CT [56]).

Ethical

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2019.105264

References

[1] W. H. Organization, International Statistical Classification of Diseases and Related Health Problems, 1, World Health Organization, 2004.
[2] H.T. Madabushi, M. Lee, High accuracy rule-based question classification using question syntax and semantics, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1220–1230.
[3] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174.
[4] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2018, pp. 328–339.
[5] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org
[6] Y. Goldberg, A primer on neural network models for natural language processing, J. Artif. Intell. Res. 57 (2016) 345–420.
[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[8] K. Bhatia, H. Jain, P. Kar, M. Varma, P. Jain, Sparse local embeddings for extreme multi-label classification, in: Advances in Neural Information Processing Systems, 2015, pp. 730–738.
[9] H. Jain, Y. Prabhu, M. Varma, Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 935–944.

[10] K. Jasinska, K. Dembczynski, R. Busa-Fekete, K. Pfannschmidt, T. Klerx, E. Hullermeier, Extreme F-measure maximization using sparse probability estimates, in: International Conference on Machine Learning, 2016, pp. 1435–1444.

[11] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, X. Geng, Binary relevance for multi-label learning: an overview, Front. Comput. Sci. 12 (2) (2018) 191–202.

[12] P. Franz, A. Zaiss, S. Schulz, U. Hahn, R. Klar, Automated coding of diagnoses–three methods compared., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2000, p. 250.

[13] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, J. Am. Med. Inform. Assoc. 21 (2) (2013) 231–237.

[14] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952.

[15] J. Pérez, A. Pérez, A. Casillas, K. Gojenola, Cardiology record multi-label classification using latent dirichlet allocation, Comput. Methods Programs Biomed. 164 (2018) 111–119.

[16] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: European Conference on Machine Learning, Springer, 1998, pp. 137–142.

[17] A. McCallum, K. Nigam, et al., A comparison of event models for naive Bayes text classification, in: AAAI-98 Workshop on Learning for Text Categorization, 752, Citeseer, 1998, pp. 41–48.

[18] J. Nam, J. Kim, E.L. Mencía, I. Gurevych, J. Fürnkranz, Large-scale multi-label text classification revisiting neural networks, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 437–452.

[19] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, doi:10.3115/v1/D14-1181.

[20] D. Tang, B. Qin, X. Feng, T. Liu, Target-dependent sentiment classification with long short term memory, CoRR, abs/1512.01100(2015).

[21] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 3485–3495.

[22] W. Yin, H. Schütze, Learning word meta-embeddings, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1351–1360, doi:10.18653/v1/P16-1128.

[23] J. Coates, D. Bollegala, Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 194–198, doi:10.18653/v1/N18-2031.

[24] O. Melamud, J. Goldberger, I. Dagan, Context2Vec: learning generic context embedding with bidirectional LSTM, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 51–61.

[25] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237, doi:10.18653/v1/N18-1202.

[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805(2018).

[27] P. Nigam, Applying deep learning to ICD-9 multi-label classification from medical records, 2016.

[28] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.

[29] H. Suominen, L. Kelly, L. Goeuriot, A. Névéol, L. Ramadier, A. Robert, E. Kanoulas, R. Spijker, L. Azzopardi, D. Li, et al., Overview of the CLEF eHealth evaluation lab 2018, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018, pp. 286–301.

[30] A. Atutxa, A. Casillas, N. Ezeiza, V. Fresno, I. Goenaga, K. Gojenola, R. Martínez, M.O. Anchordoqui, O. Perez-de Viñaspre, IxaMed at CLEF eHealth 2018 task 1: ICD10 coding with a sequence-to-sequence approach, in: CLEF (Working Notes), 2018, p. 1.

[31] M. Almagro, S. Montalvo, A.D. de Ilarraza, A. Pérez, MAMTRA-MED at CLEF eHealth 2018: a combination of information retrieval techniques and neural networks for ICD-10 coding of death certificates, in: CLEF (Working Notes), 2018, p. 1.

[32] M. Oronoz, K. Gojenola, A. Pérez, A.D. de Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in Spanish: mining adverse drug reactions, J. Biomed. Inform. 56 (2015) 318–332.

[33] M. Marimon, B. Fisas, N. Bel, J. Vivaldi, S. Torner, M. Lorente, S. Vázquez, M. Villegas, The IULA Treebank, in: Lrec, 2012, pp. 1920–1926.

[34] A. Duque, M. Stevenson, J. Martinez-Romo, L. Araujo, Co-occurrence graphs for word sense disambiguation in the biomedical domain, Artif. Intell. Med. 87 (2018) 9–19.

[35] S.M. Jiménez-Zafra, M. Taulé, M.T. Martín-Valdivia, L.A. Ureña-López, M.A. Martí, SFU review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns, Lang. Resour. Eval. 52 (2) (2018) 533–569.

[36] S. Santiso, A. Pérez, A. Casillas, Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery, IEEE J. Biomed. Health Inform. 23 (5) (2019) 2148–2155.

[37] M. Almagro, R. Martínez Unanue, V. Fresno Fernández, S. Montalvo Herranz, Estudio preliminar de la anotación automática de códigos CIE-10 en informes de alta hospitalarios, SEPLN, 2018.

[38] H. Fabregat, A. Duque, J. Martinez-Romo, L. Araujo, Extending a deep learning approach for negation cues detection in Spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain, 2019, p. 1.

[39] H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014, p. 1.

[40] Y.-T. Zhou, R. Chellappa, Computation of optical flow using a neural network, in: IEEE International Conference on Neural Networks, 1998, 1988, pp. 71–78.

[41] J. Liu, W.-C. Chang, Y. Wu, Y. Yang, Deep learning for extreme multi-label text classification, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, pp. 115–124.

[42] Y. Yin, Y. Song, M. Zhang, Nnembs at semeval-2017 task 4: neural twitter sentiment classification: a simple ensemble method with different embeddings, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 621–625.

[43] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146, doi:10.1162/tacl_a_00051.

[44] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[45] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[46] C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2016.

[47] K. Lee, L. He, L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 687–692, doi:10.18653/v1/N18-2108.

[48] M. Fares, A. Kutuzov, S. Oepen, E. Velldal, Word vectors, reuse, and replicability: Towards a community repository of large-text resources, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 271–276.

[49] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, N. Taright, Supervised topic models for diagnosis code assignment to discharge summaries, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 485–497.

[50] C. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Nat. Lang. Eng. 16 (1) (2010) 100–103.

[51] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, 2013, pp. 1310–1318.

[52] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: Advances in Neural Information Processing Systems, 2017, pp. 3856–3866.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[54] Y. Zhang, J. Schneider, Multi-label output codes using canonical correlation analysis, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 873–882.

[55] J. Wehrmann, R. Cerri, R. Barros, Hierarchical multi-label classification networks, in: International Conference on Machine Learning, 2018, pp. 5225–5234.

[56] K. Donnelly, SNOMED-CT: the advanced terminology and coding system for ehealth, Stud. Health Technol. Inform. 121 (2006) 279.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 30, 2020, accepted October 3, 2020, date of publication October 7, 2020, date of current version October 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029429

Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time

ALBERTO BLANCO, ALICIA PÉREZ, AND ARANTZA CASILLAS
HiTZ Center-Ixa, University of the Basque Country (UPV/EHU), 20080 Donostia, Spain

Corresponding author: Alberto Blanco (alberto.blanco@ehu.eus)

This work was supported in part by the Spanish Ministry of Science and Technology under Grant PAT-MED PID2019-106942RB-C31, in part by the Basque Government under Grant Elkartek KK-2019/00045 and Grant IXA IT-1343-19, and in part by the Predoctoral under Grant PRE-2019-1-0158.

ABSTRACT This work deals with clinical text mining for automatic classification of Electronic Health Records (EHRs) with respect to the International Classification of Diseases (ICD). ICD is the international standard for the identification of diseases and health conditions in EHRs and the foundation for reporting health statistics. Machine learning-based techniques have proven robust to infer classification models from EHRs. Since each EHR tends to involve multiple diseases, multi-label classification is required. The concern in this work is the versatility of the models inferred and their ability to generalise in two ways: as time goes ahead and across hospital services or health specialties. Indeed, in this work, we show the capabilities of a Bidirectional Recurrent Neural Network (RNN) with GRU units and ELMo embeddings on two corpora (a corpus comprising a set of EHRs within the Basque Health System, namely Osakidetza, and the well-known MIMIC-III corpus). To delve into and assess the versatility of the models, we focus on their resilience across hospital admissions taken over two different years and also across six distinct hospital services. In addition, we paid attention to the classification performance to estimate ICD codes of different granularity (e.g. with or without essential modifiers). Our best results are 39.55% and 47.28% F-Score for the Osakidetza and MIMIC-III datasets respectively, with the original main label-sets. Regarding the models evaluated per specialty, the most remarkable results are 57.00% and 72.74% F-Score, in the Cardiology and Nephrology medical services respectively.

INDEX TERMS Extreme multi-label classification, electronic health records, international classification of diseases, classification across-time, classification across hospital-services.

I. INTRODUCTION

Natural Language Processing (NLP) is gaining relevance within the clinical documentation services to cope with extensive information conveyed by Electronic Health Records (EHRs). Healthcare data is getting increasingly larger and complex to process [1], but evidence shows its usefulness in such different sectors as Adverse Drug Reaction extraction [2], [3] and identification of complex symptoms, assessed in several cohorts of patients in hemodialysis [4], as well as relevant symptoms in patients with schizophrenia [5], and breast cancer [6], and the creation of phenotypes to characterise patients [7], [8].

Facilitating access to information is crucial for accurate clinical documentation. International Classification of Diseases (ICD) [9] is a standard used to classify diagnosis and procedures within EHRs. These codes are used to quantify vital statistics, for surveillance, to seek cohorts of patients with similar diagnoses in downstream studies and also as a standardised information exchange method between hospitals. The thorough and accurate coding of EHRs affects critical clinical information extraction and also other industries such as insurance billing [10]–[12].

Nowadays EHRs are manually encoded by healthcare professionals specially trained to cope with complex ICD nuances. Note that ICD-10 is arranged in 24 chapters or branches of medicine and comprises nearly 70 thousand codes for diseases (ICD-10-CM) and as many for procedures (ICD-10-PCS). The ICD versions evolve rapidly e.g. from the 9th version to the 10th the number of codes increased five-fold and, what is more, the code structure changed from a maximum of 5 characters to 7; the alpha-numeric coding

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey.

structure was also modified. The chapter is encoded in this structure and would reflect a coarse-grained classification of the EHR into branches referred to as *ICD chapter*. Besides, in this coding we typically find a sub-set of characters (often 3 of them) that are referred to as the *main ICD class*, and the remaining characters comprise what is referred to as non-essential modifiers (e.g. laterality and severity). Altogether these are referred to as '*fully-specified ICD class*'. Additionally, manual coding is time-consuming. As an example, let us focus on a well-known collection of EHRs, i.e. MIMIC-III [13], a set of nearly 55, 000 EHRs made available by the MIT Lab for Computational Physiology from the Beth Israel Deaconess Medical Centre. Each EHR has, on average, 1, 947 words that are carefully read by professionals to annotate the EHR with all the ICDs found. On average, each of these EHRs was assigned 11.5 different ICDs (also referred to as *label cardinality*) and, altogether, the set shows 6, 527 different ICDs (also known as the size of the label-set). Given that manually encoding is time-consuming and requires specialised professionals, several health systems took the decision to restrict coding to just the main cause of admission, leading to a relevant loss of valuable information. It is well-known that EHRs show much variety in terms of (often non-standard) linguistic forms, semantics and syntax, and a linguistic style prone to the economy of language [14], [15].

There is evidence that NLP can aid human coders in a decision support system with the human in the loop [16]. The task of automatically assigning multiple codes to a given document is referred to as *multi-label classification*. Automatic multi-label classification of EHRs is, however, far from the scope of current machine learning approaches given that high accuracy is a must. There are numerous **challenges** inherent in this task, not only from the inference perspective but also due to the assessment of the quality of the predicted label-set. The size of the output label-set (e.g. 6, 527 in MIMIC-III) is extremely high for the input evidence (e.g. 55, 000 EHRs in MIMIC-III). While automatic inference can find patterns, pattern repetition in selecting an unknown number (11.6 labels on average in MIMIC-III, with 9.0 as the median value and a standard deviation of 6.3) of ICD codes from a set of 6, 527 lead to 51,980 unique label-sets. Note that there are few pieces of evidence for the output on top of the variability in the input. This highly-variable classification task with thousands of possible labels is referred to as *extreme multi-label classification (XMC)* [17]. Quantitative assessment of multi-label models is still one of the stumbling blocks in the inference since model optimisation (fine-tuning) rests on evaluation approaches and this is not a trivial issue. (We shall discuss issues that might emerge in the assessment in Section III-B).

In this work, we assess a multi-label classification approach in the task of EHR multi-class classification in documents written in Spanish. The set of EHRs is comparable in many aspects to those in MIMIC-III (as we will present in Section IV-A). Often, [18], [19] related work bound the size of the label-set in such a way that the data-set ensures a minimum number of documents per label (in an attempt to ensure minimum repetition per pattern). In this work we assessed the system exhaustively.

First, to assess the resiliency of the learning approach proposed, i.e. Bidirectional Recurrent Neural Network [20] with GRU units [21] and ELMo embeddings [22] (BiGru ELMo), the label-set was not restricted and all the labels available in the data (i.e. 2, 554 labels) were considered and next, the system was assessed with the top 110 and top 16. Second, we assessed the robustness of the system across time. Needless to say, as time goes by, personnel in a hospital might have changed their EHR writing or encoding style: in these circumstances, the system should be adapted. The question also arises here of how often should we re-train the model and also whether previous EHRs are either beneficial or harmful for current EHR coding. The motivation is to assess whether a predictive model inferred with data from a given year can help to predict EHRs from future years. Also, we want to evaluate if non-overlapping data from two consecutive years help to predict EHRs from the later year of the two. In other words, we wish to assess/evaluate if training with data from successive years can generate synergies or, whether the best option is to re-train the system frequently to keep it updated. Third, we did not only assess the system across time but also across use-cases in different hospital services. One of our concerns had to do with data scarcity. We wondered if a general system trained with EHRs from discharge reports from several hospital services (e.g. cardiology, psychiatry etc.) is able to cooperate and make the system capture accurate syntax and semantic nuances, or if, by contrast, accurately encoding EHRs from a given service was bound to train the system with EHRs from that service, while EHRs from other services could lead to lexical explosion and maybe distort the outcome. Through these experiments we tried to shed light on the following three **research questions**: 1) the ability of BiGru ELMo to cope with infrequent and frequent labels, 2) the robustness of the model across time and, 3) across hospital services. Briefly, the novelty of the work resides in the aforementioned research questions. To this end, we apply a state-of-the-art multi-label classification model to a Spanish EHR dataset that can be segmented by year and medical service. The segmentation of data allows checking the robustness of the models against lexical variation due to variations among medical specialties and across-time. We also assessed the model in both coarse-grained (the ability of the model to situate the EHR within a chapter of the ICD) and fine-grained code assessment (referring to the granularity mentioned on page 183534). The finer the granularity, the bigger the size of the label-set.

II. RELATED WORK

Since 2000 CLEF has organised different laboratories in the field of multilingual access evaluation, in particular since 2016 in the automatic assign of ICD codes. In 2016 [23] the task consisted of extracting causes of death from French narratives as coded in the International Classification of Diseases

ICD-10. In 2017 [24] the task goal was to automatically assign ICD-10 codes to English and French death certificates. In 2018 [18] the task focused on French, Hungarian and Italian texts. In 2019 [25] the task explored the automatic assignment of ICD-10 codes to non-technical summaries of animal experimentation in German. The tasks carried out from 2016 - 2018 are focused on the codification of lines (diagnosis) instead of on the codification of whole EHRs. On average, each diagnosis has between 2.06 to 12.38 tokens and between 1.20 to 1.37 codes.

Approaches based on regular expressions or transducers either manually created [26], [27] or automatically inferred from data [28] were used in previous works when it comes to mapping Diagnostic Terms (DT) expressed in natural language into standard DTs within the ICD and, hence, assigning the corresponding ICD. The difference between translating non-standard expressions to a standard form and assigning ICDs to a given full EHR is substantial. The entire EHR in our task has on average ∼1,000 words per document, while the input non-standard DT tends to have around 5 words. In the EHR, the language is likewise, non-standard, although, the DTs are not explicitly informed. Moreover, implicit evidence, such as analytics and current treatment, might yield an ICD. Besides, while the correspondence between the non-standard DT and the ICD codes is 1 to 1, in the EHRs, 1 short phrase could trigger n ICD codes being the correspondences m to n.

The so-called *binary relevance approach* [29] is a simple approach to tackle multi-label classification that comprises as many binary classifiers as classes involved. Each classifier would determine the absence or presence of one class. The drawback of this simplistic approach is that the classes are assumed to be independent, hence, dependencies among ICD codes would be disregarded. Nevertheless, some diagnostics are incompatible (and should not be predicted together), while others tend to co-occur. Accordingly, we opted for a model that considers the label-dependencies.

Rios and Kavuluru [30] explored the use of Convolutional Neural Networks (CNNs) for automatic ICD coding. They stated that when many codes occur infrequently, the Deep Learning (DL) models' performance is inhibited. They proposed a neural transfer learning strategy, supplementing EHR data with PubMed indexed biomedical research abstracts. For the source task, they trained a CNN to predict 1.6M Medical Subject Headings (MeSH) using PubMed indexed biomedical abstracts, whereas, for the target task, they trained a CNN on 71,463 EHRs to predict ICD diagnosis codes. Our approach is also based on the idea of transfer learning, as the ELMo embeddings are derived from a bidirectional LSTM trained with a coupled language model (LM) objective on a large text corpus, including, but not restricted to biomedical texts (i.e. pharmaceutical or medical articles from Wikipedia). They got, respectively, a micro- and macro-F-Score of 56.8 and 28.6, considering 1, 231 truncated ICD-9 labels with 5, 303 average words per instance from 71, 463 instances.

Gangavarapu *et al.* [19] employed the MIMIC data-set. It is usual to exploit the discharge summaries (i.e. the clinical report prepared by the physician after a hospital stay), but in this case, they leverage the nursing notes. One drawback is that the nursing notes present excessive redundant information, due to the anomalous and evolutive data of the patient. This issue was addressed with a fuzzy similarity-based data cleansing approach; The authors applied vector space and topic modelling to extract the rich patient-specific information available in unstructured clinical data. This can be crucial in countries where structured EHR adoption is not widespread [31], [32]. The authors worked with 223, 556 nursing notes of 357.8 words on average, predicting 19 ICD-9 code group labels, and achieved a maximum F1-score (weighted-) of 69.81 across all the tested models.

Most ICD codes appear only in a few samples, that is, the ICD distribution presents a long-tail, which is, precisely, a feature of extreme multi-label classification. Babbar and Schölkopf [33] posed the tail-label detection task in XMC as a robust learning problem, taking into account the worst-case perturbation scenarios. This viewpoint is motivated by a key observation: from the training set to test set, there is a significant change in the distribution of the features of instances belonging to the tail-labels. This is a typical case when classifying EHRs with ICD codes, especially, across time or clinical services [34] since physicians from different medical specialties refer to the same medical concepts in diverse forms. The converse also happens: the same string is employed to refer to different concepts (this happens often with abbreviations) across clinical services.

In an attempt to tackle the scalability issue of state-of-the-art Deep Learning-based methods to extremely large label-sets [35], a hierarchical structure based on Probability Label Trees generated with balanced k-means recursively, and multi-label attention was proposed by You *et al.* [36]. Similarly, Gargiulo *et al.* [37] presented a methodology named Hierarchical Label-Set Expansion (HLSE), used to regularise the data labels, based on the hierarchical structure of the MeSH label-set. Data scarceness and large lexical variability and vocabularies are major concerns in the ICD multi-label classification tasks. Deng *et al.* [38] presented a processing pipeline built upon CNNs and an autoencoder with logistic regression. They applied the combination of embeddings from different sources and proved the positive influence of semantic enrichment to counter the aforementioned strains. The contextual ELMo embeddings can overcome these limitations of the standard embeddings [39]. Cheng *et al.* [40] recognised that some complex semantic problems in the real world require the association of more objects with related labels but also that as data complexity increases, the class imbalance issues become increasingly prominent. A well-known strategy to deal with imbalance between classes is to use label correlations, but their work proposed an alternative approach. They first introduced the classification margin and expanded the original label-space among labels, taking into account the label-density. The BiGru model can handle all

the labels at once thanks to the final dense layer with the Sigmoid activation function, and thus, is able to capture and model the label dependencies. Chalkidis *et al.* [41] compared various neural methods on the EURLEX57K dataset (with 4,271 labels) and concluded that the best results rely on the Recurrent Neural Network with GRU units, but also that it is the most computationally expensive method. Chang *et al.* [42] leveraged the pre-trained language representation model BERT, extending it to the XMC problem to deal with the difficulty of capturing dependencies or correlations among labels and the tractability to scale to the extreme label setting because of the Softmax bottleneck scaling linearly with the output space. Their so-called X-BERT utilises both the label and input text to build label representations. This induces semantic label clusters to better model label dependencies, which can also be applied to the ICD classification task, as all the labels have an associated text, the standard-term description.

The **motivation** and novelty of this work resides in exploring the behaviour of the classifiers under novel circumstances through the characteristics of our task. It conveys a multi-label text classification problem with great lexical variability, especially in the set of EHRs in Spanish, as the dataset can be segmented by year and medical service. To that end, and following the insights of the related works, we developed a model based on Recurrent Neural Networks with Bidirectional Recurrent layers and GRU units leveraging the ELMo contextual embeddings. These models were proven robust and capable of learning from scarce samples, as is the case of ICD coding. The gaps found in previous works, and which we do cover in this article, are related to the capability of such state-of-the-art models to keep a strong performance facing lexical variation inherent to the biomedical domain, but extended to variations over time (i.e. attempting to make predictions across years) and different sub-domains (i.e. across various clinical services). This way, we assess the sensitivity of these models to different factors (time and health services) and, thus, pay attention to their usability.

III. METHODS

A. MULTI-LABEL CLASSIFICATION APPROACH

Having explored previous works, for our task we opted for a Recurrent Neural Network with a Bidirectional layer with GRU units (referred to as BiGru from now onward). The architecture of the model is shown in Figure 1, and formally, is explained in (1), with the bidirectional layer processing the sequences of text in both directions, forward and reverse. Accordingly, it generates forward ($\overrightarrow{h^{(t)}}$) and backward ($\overleftarrow{h^{(t)}}$) hidden states, which are later combined into $h^{(t)}$. Here t is the time-step and T the total number of time-steps ($1 \leq t \leq T$). The parameters to be determined in the inference stage given the EHRs are, on the one hand, the weight matrices, W and V, and, on the other hand, the bias term b. A non-linear activation function, the Sigmoid (σ), is chosen to compute the current hidden-states taking, as input, the weighted sum of previous

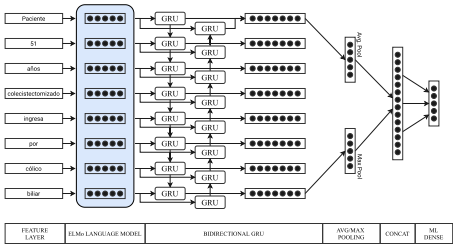


FIGURE 1. Architecture of the BiGru ELMo model: a Bidirectional Recurrent Layer with GRU units powered by ELMo embeddings with Pooling and a final multi-label dense layer.

hidden-states ($h^{(t-1)}$) and current input ($x^{(t)}$) with the weights given by W and V. Finally, both hidden states are combined as a mere concatenation of each matrix.

$$\begin{aligned}\overrightarrow{h^{(t)}} &= \sigma(\overrightarrow{W}x^{(t)} + \overrightarrow{V}\,\overrightarrow{h}^{(t-1)} + \overrightarrow{b}) \\ \overleftarrow{h^{(t)}} &= \sigma(\overleftarrow{W}x^{(t)} + \overleftarrow{V}\,\overleftarrow{h}^{(t-1)} + \overleftarrow{b}) \\ h^{(t)} &= [\overrightarrow{h}^{(t)} \| \overleftarrow{h}^{(t)}]\end{aligned} \tag{1}$$

The output of the bidirectional layer, $h^{(t)}$, is fed to several Pooling [43] layers. To be precise, in our work, an average and a max-pooling layer were used, as in (2).

$$h_{max} = \max_{1 \leq t \leq T} h^{(t)} \text{ and } h_{avg} = \frac{\sum_{t=1}^{T} h^{(t)}}{\left\| h^{(t)} \right\|} \tag{2}$$

The output of both pooling layers (h_{max} and h_{avg}) is again concatenated into $h = [h_{max} \| h_{avg}] \in \mathbb{R}^{2T}$ and passed into a final dense layer, which is responsible of computing the probability estimation of the labels i.e. ICD codes.

The strength of RNN with GRU unit and ELMo is clear in this extreme multi-label scenario as described in what follows. This BiGru model can cope with the multi-label problem since the final dense layer is able to capture and model the label dependencies in contrast with the binary-relevance approach. In fact, by virtue of the Sigmoid activation function, it produces a probability estimation for each label that is not mutually exclusive [17]. Thus, this BiGru model is suitable to cope with the multi-label classification of EHRs through ICDs and address dependencies between diagnoses. Moreover, bearing in mind that EHRs are long documents, with even thousands of words per document, the ability to capture long-term dependencies in the text, as an RNN with GRU does, becomes imperative. Furthermore, in an attempt to cope with lexical variability in the input EHRs, we turned to ELMo embeddings. In general, the word embedding is a technique to transform a word, and therefore a document, into a dense vector and, by extension, into a matrix. The text from a clinical record is fed into an embedding layer, and the output is a matrix representing the document, with one row-vector per word. Each word is referred to as time-step in the formulation of the model. ELMo embeddings [22]

capture contextual information and, according to the authors, these word representations model i) complex characteristics of word use (i.e. syntax and semantics) and ii) how these uses vary across linguistic contexts (i.e. to model polysemy). As a result, in our task, ELMo embeddings can help towards i) the detection of some essential nuances of medical records such as the negation of symptoms and, ii) robustness to variations of the author (i.e. each physician expresses in a particular way) and sub-domain (i.e. a reference to the same medical concepts across several clinical services). We also need embeddings, as they can cope with different linguistic contexts, to deal with the various clinical services, time-frames, and their lexical subtleties. These are the principal reasons why we opted for BiGru ELMo as an appropriate choice to deal with high lexical variability within EHRs and multi-label classification with respect to the ICD.

B. MULTI-LABEL ASSESSMENT CRITERIA

We have resorted to well-known metrics such as Precision, Recall and F-Score, but to adapt them correctly to the multi-label scenario it is necessary to compute averages since these metrics are well suited to mono-label tasks. There are several common averages, i.e. micro-, macro- and weighted-average-, each of them penalising certain aspects more severely than other [44], [45]. For example, a macro-average will compute the metric independently for each class (i.e. a confusion matrix for each ICD) and then compute each metric and take the average of the metric (hence it will treat all classes equally). By contrast, a micro-average will aggregate the contributions of all classes in a single confusion matrix and then compute the average (hence, the performance over more populated classes dominate). Thus, in a situation with imbalanced classes is easier to get higher metric values with the micro-average provided that the dominant class is accurately predicted (with the micro-avg result being almost insensitive to the hits or fails over less populated classes). The weighted-average is a balanced solution which takes into account the support (i.e. frequency) of each label to weight their contribution to the final metric value. For that reason, in this work, we give the weighted-average version of the Precision, Recall and F-Score metrics. Nevertheless, all these approximations and variations come with benefits and disadvantages: there is not a general optimum approach, and the best-suited evaluation will depend on the task and objectives of the work. Note, however, that averages are taken per code and not, strictly, per document. A challenge in ML classification is to on decide the number of codes to assign to each document, as this is variable (on average, EHRs within MIMIC receive 11.6 codes but the deviation is 6.3, quite high).

IV. EXPERIMENTAL FRAMEWORK

A. CORPORA

Here we describe the corpora and provide two perspectives: the input (text from EHRs) and the output (ICDs). The data we had available for this work comprised two separate but analogous data-sets with EHRs, written in Spanish, from the Basque public health system (Osakidetza). Specifically, both sets, denoted as Osa1 and Osa2, comprise discharge summaries from hospitals. Table 1 provides quantitative details of each data-set and the union of both sets (denoted as Osa1+2) leading to a total of 27, 040 unique EHRs.

TABLE 1. Quantitative description of the input (EHRs).

	data-set			
	MIMIC	Osa1	Osa2	Osa1+2
Samples	55,172	13,574	13,466	27,040
Vocab. size	137,207	236,109	255,394	379,477
Words/doc	$1{,}399 \pm 721$	843 ± 401	885 ± 428	864 ± 415

Regarding the input (EHRs), both Osa1 and Osa2 data-sets are significantly smaller than MIMIC; indeed, there are nearly twice as many samples in MIMIC as in Osa1+2. However, the size of the vocabulary (the number of unique words) of Osa1+2 (379, 477) is approximately three times larger than the vocabulary of MIMIC (137, 207). This means that the lexical variability is notably higher in the Spanish set, even though there are more documents and, besides, the length of the documents is much higher in MIMIC. To enable the drawing of conclusions from the following experiments, we must acknowledge the distributions of the features of the different sets from the experimental setup. To that end, as the classifiers are only fed with the text from the clinical records, we explore the vocabulary and Out-of-Vocabulary (OOV) words. The number of unique words in Osa1 is 89,840, the number of unique words in Osa2 is 94,764, and the number of OOV words in Osa2 with respect to Osa1 is 42,249, which is the 44.6% of the vocabulary. The vocabulary we are dealing with is large, but what is more, we can observe that the amount of disjoint vocabulary between sets is also quite high, leading to the demand for robust classifiers.

Regarding the output, i.e. the **label-sets**, the aim is to predict the set of ICDs in their fully-specified form. However, failing non-essential modifiers might be considered not as bad as failing the chapter of the ICD. Accordingly, we assessed the performance taking into account three different granularities of the ICD codes, from fine- to coarse-grained:

- *"Full-code"* level preserves the original code e.g. "`I13.10`" (this stands for *Hypertensive heart and chronic kidney disease without heart failure, with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease*);
- *"Main"* level drops non-essential modifiers, keeping just the first three characters e.g. "`I13`" (*Hypertensive heart and chronic kidney disease*)
- *"Chapter"* level keeps only the first character, that is, the chapter of the ICD e.g. "`I`" (*Diseases of the circulatory system*).

Trying to predict rare codes is really challenging for inferred systems and often previous works pruned the

TABLE 2. Size of label-set ($|C|$) taking different granularity of labels for each corpus. Three levels of granularity were assessed: "Full" stands for "fully specified ICD code", "Main" for the essential modifiers, "Chapter" for the ICD chapter. "Full$_{fi\%}$" is the subset of Full in which the ICD codes were seen in at least *fi*% documents (∼1% and ∼5% of the documents respectively).

	$\|C\|$			
	MIMIC	Osa1	Osa2	Osa1+2
Full	6,918	2,554	2,554	2,554
Main	941	870	870	870
Chapter	12	24	24	24
Full$_{1\%}$	110	110	110	110
Full$_{5\%}$	16	16	16	16

label-set according to a minimum frequency threshold [46], [47]. In an attempt to make comparisons with respect to previous works, we created a sub-set of instances restricted by frequency. $|C_{train_r}|$ denotes the size of the label-set restricted to most prevalent ICDs following repetition boundaries shown in previous works. Note that, $|C_{train_r}| = |C_{train}|$ means that no restriction was applied. We experimented with a subset from the full label-set in which the occurrences of the codes were above a threshold f. In our case, we considered two thresholds leading to two label-sets, denoted as Full$_{1\%}$ and Full$_{5\%}$, which incorporated a code whenever it appeared in at least ∼1% and ∼5% of the documents respectively. Quantitative details of the label-set in our data-sets are given in Table 2. The table reveals substantial sparsity: just 110 out of 2, 554 ICDs appear at least in 1% of the EHRs (i.e. diseases diagnosed around a hundred times in a set of around 10, 000 EHRs). Note that for all the experiments carried out, the train and test partitions are obtained with the iterative stratification algorithm, with a 70/30 split.

B. PERFORMANCE BY LABEL GRANULARITY

Table 3 shows the experimental results of the model on each corpus (Osa1, Osa2, Osa1+2, and MIMIC). Regarding the label-set reduction, the assessment was made on each of the aforementioned label sub-sets (Full, Full$_{1\%}$, and Full$_{5\%}$ presented in Table 2). Additionally, Table 3 shows the assessment of the computer-aided coding system to various levels of granularity.

Note that increasing the size of the label-set from 110 to 2, 500 (an increment of 1 to 22), as we could expect, was detrimental to the F-Score (from 37.87 to 20.43). Nevertheless, the decrease was not as dramatic as 22 to 1 and the same applies to the results of just 16 prevalent labels. This insight suggests that the model is able to learn from rare cases in EHRs (as one-shot learning strategy aims to) and is able to make predictions of ICDs with little prevalence.

Experiments carried out with the entire label-set (with above 2, 550 different ICDs) show reasonable performance in terms of averaged scores. Although it is difficult to make a fair comparison because no standardised set of experiments has been popularised on any ICD code-based multi-label classification data set, there are some reference works with which we can validate the performance of our models.

TABLE 3. Performance of the system over different ICD code-lengths or granularity (F: Full, M: Main, C: Chap) for all specialties together. P denotes Precision, R Recall and F the F-Score.

Train	Eval	EHRs	grain	$\|C_{train_r}\|$	P	R	F
Osa1	Osa1	13,574	F	2,553	30.61	16.77	20.43
		11,370		110	53.41	31.54	37.87
		8,732		16	70.97	50.44	57.08
		13,574	M	870	44.59	34.36	37.50
		9,844		19	76.10	59.98	66.31
		13,574	C	24	78.42	64.54	69.90
		13,513		14	82.00	66.26	72.51
Osa2	Osa2	13,466	F	2,553	24.49	15.93	18.56
		11,635		110	48.78	26.85	33.02
		9,323		16	67.80	45.99	53.77
		13,466	M	870	40.28	32.28	34.99
		10,142		19	71.86	58.06	63.73
		13,466	C	24	75.46	61.64	66.59
		13,392		14	78.82	66.10	68.27
Osa1+2	Osa1+2	27,040	F	2,554	27.64	20.52	22.48
		22,907		110	53.30	33.79	39.89
		18,098		16	70.14	53.52	59.84
		27,040	M	870	44.97	36.92	39.55
		19,925		19	73.56	63.46	67.52
		27,040	C	24	79.96	66.18	71.77
		26,905		14	80.88	67.46	72.85
MIMIC	MIMIC	55,172	F	6,902	30.72	34.87	31.43
		55,172		110	50.29	53.12	51.26
		55,172		16	67.40	63.81	65.42
		53,090	M	941	46.75	49.55	47.28
		45,189		19	69.47	66.61	67.96
		48,120	C	12	79.12	77.85	78.44

Dermouche *et al.* [48] obtained 75.0 micro F-Score and 35.0 macro F-Score with a Support Vector Machine (SVM) model, and 74.0 micro F-Score and 38.0 macro F-Score with a Latent Dirichlet Allocation (LDA) model, but taking into account just 252 codes from the MIMIC dataset, and computing the F-Score retrieving the correct class among the 10 most probable classes returned by the model. Duarte *et al.* [47] achieved 27.04 macro F-Score with their best model based on Hierarchical GRUs considering the Full codes, 40.50 considering main codes and 62.91 considering chapter codes. The number of labels was 1,418, 611 and 19 respectively. In brief, taking as the baseline the aforementioned state of the art approaches, our approach is ahead in several aspects. In light of the results attained with MIMIC-III, the model was proven competitive with respect to previous works in a fully automatic classification scenario entailing fully-specified ICD codes. Having validated the results on a well-known corpus, we have extended the study to the corpus from Osakidetza (a set of EHRs in Spanish).

Our system was assessed with two non-overlapping sets of EHRs from two different years from Osakidetza, named Osa1 and Osa2. As shown in Tables 1 and 2, both Osa1 and Osa2 have a similar number of input EHR texts (about 13, 500 documents). The size of the label-set is the same (2, 550 in round figures), as is the label cardinality (on average, 5.8 labels per document). However, the F-Score differs 2 absolute points in the most extreme case with all the labels (from 20.43 to 18.56). Nevertheless, as expected due to

TABLE 4. Behaviour of current model on unseen current and future EHRs, for all the specialties together, and with granularity Full. P denotes Precision, R Recall and F the F-score.

Train	Test	EHRs	$\|C_{train}\|$	$\|C_{train_r}\|$	$\|C_{test}\|$	P	R	F
Osa1	Osa1	13,574	2,553	2,553	2,023	30.61	16.77	20.43
Osa1	Osa2	13,574	2,553	2,553	2,052	21.10	8.95	11.61
Osa2	Osa2	13,466	2,553	2,553	2,052	24.49	15.93	18.56
Osa1+2	Osa2	27,040	2,554	2,553	2,023	26.22	19.36	21.52
Osa1+2	Osa1+2	27,040	2,554	2,554	2,554	27.64	20.52	22.48

TABLE 5. Osa1+2 generalist model trained on the $\text{Full}_{1\%}$ label-set but re-evaluated per specialty and also applying the 'specialty labels' label-set modification.

Specialty	EHRs	$\|C_{train}\|$	$\|C_{train_r}\|$	$\|C_{test}\|$	C_{modif}	P	R	F
Pneumology	5,199	2,057	110	107	All	50.44	37.17	40.92
				17	Spec	55.62	45.33	48.39
Cardiology	4,731	1,470	110	105	All	54.00	39.84	43.66
				23	Spec	61.53	49.61	53.22
Digestive	3,990	1,896	110	108	All	50.46	25.80	32.39
				6	Spec	27.17	20.12	23.02
Psychiatry	1,724	697	110	46	All	43.25	30.33	28.95
				7	Spec	54.75	47.06	42.14
Hematology	1,350	1,088	110	98	All	55.77	28.77	36.76
				4	Spec	56.52	20.29	29.86
Nephrology	841	785	110	94	All	59.97	39.87	46.43
				7	Spec	82.95	64.88	72.74

the higher number of available samples, the best performance is attained with the union of both data-sets, even though the union conveys ∼500 labels more on the test set. With the union, the results from Osa1 are improved by another two points, leading to an F-Score of 22.48 on the full label-set.

Bearing a computer-aided ICD classification system in mind, we assessed the model in three scenarios with increasing details in the predicted label, ranging from the ability of the model to predict the fully-specified ICD code (denoted as F (Full) in Table 3), the main class without non-essential modifiers (denoted as M (Main) in Table 3) or the ICD chapter (denoted as C (Chap) in Table 3). Given an EHR, the model is shown to be effective in the assignment of the chapters of the ICD, restricting the use to just those chapters, which can be useful, as discussed in Section V. As the label gets more and more specific, the task gets more complex. Restricting the granularity of the model impacts upon the size of the label-set: while there are thousands of fully-specified ICDs, there are just 24 chapters and 870 main labels. Note that with the Osa1+2 data-set, the F-Score with the Full granularity and the label-set reduced from 2,554 to 110 labels is 39.89, almost the same as the 39.55 F-Score obtained with the non-reduced 870 Main labels.

C. SENSITIVITY OF THE MODEL TO LEXICAL VARIANTS ACROSS TIME

Would a system learn consistently from EHRs issued in one year how to classify EHRs in the future? How much does data addition boost the performance of the system? These experiments would suggest that the lexical features within EHRs (possibly clinical personal, clinical specialisations, etc.) changed over time. Also, note that the number of samples can influence the results, especially when the concatenation of both data-sets are used for the training of the model.

Table 4 shows the ability of each model to classify current and forthcoming EHRs. To this end, the model was trained with current and past EHRs and assessed with either current events or events from subsequent years unclassified at that moment. The aim is to test the sensitivity to different time-frames. As we could expect, training the models with EHRs issued in the same year as those in the evaluation is beneficial, and even more so, if the training set is completed with EHRs even from previous years. As we can see in the rows with Test = Osa2 when the training data is from both years, the F-Score raises from 18.56 to 22.48, increasing the performance by ∼4 points, and it is ∼2 times higher than when training only with EHRs from a previous year (11.61 to 22.48).

D. SENSITIVITY OF THE MODEL TO LEXICAL VARIANTS BY HOSPITAL SERVICE

The full data-set comprises discharge reports issued by different hospital services: e.g. Cardiology, Digestive, Neurology, etc. While decreasing the amount of EHRs tends to be detrimental to the effectiveness of the inference process, restricting the service may also reduce the lexical variability in the input and, eventually, might benefit the predictive ability. In other words, we aim to respond to the following question: how sensitive are the generalist model and the Specialty Models are to relevant codes belonging to the specialty in which the patient was admitted?. These results are shown in Tables 5 and 6.

TABLE 6. Specialty Models trained on subsets of EHRs from the Osa1+2 data-set by specialty, with the same $\text{Full}_{1\%}$ label-set and also applying the 'specialty labels' label-set modifications as in Table 5 to enable comparison.

Specialty	EHRs	$\lvert C_{train}\rvert$	$\lvert C_{train_r}\rvert$	$\lvert C_{test}\rvert$	C_{modif}	P	R	F
Pneumology	5,199	2,057	110	110	All	44.05	25.84	30.49
			17	17	Spec	61.50	41.54	45.57
Cardiology	4,731	1,470	110	110	All	50.12	32.17	37.90
			23	23	Spec	66.83	52.69	57.00
Digestive	3,990	1,896	110	110	All	64.01	14.54	21.28
			6	6	Spec	73.01	45.59	52.38
Psychiatry	1,724	697	110	110	All	30.09	24.92	25.40
			7	7	Spec	70.36	48.04	51.74
Hematology	1,350	1,088	110	110	All	66.93	26.18	35.42
			4	4	Spec	53.33	55.00	43.74
Nephrology	841	785	110	110	All	61.45	31.87	39.60
			7	7	Spec	67.56	72.27	66.81

To deal with this question, we begin with Table 5, where we present the performance of a generalist model (from Table 3, trained with EHRs from all the medical services) evaluated over the different subsets of EHRs by medical service. On the other hand, we also trained a model specifically for each service (namely, Specialty Models) limiting the training data to the EHRs issued in that service. The results are shown in Table 6.

Individual models were created, one per clinical service. Each model was trained receiving only discharge reports from a single service. While this reduces the documents accessed by each model, if the language boundaries are subject to lexical nuances particular to individual services, the models might show good performance, particularly in predicting ICD codes from the service with which they were trained. Note, however, that even though the EHRs were restricted to a single service, they convey both codes from the specialty (associated with the cause of admission in that service) as well as other codes regarding the general status of the patient and other findings (e.g. ex-smoker and type-2 diabetes).

For that reason, we consider two alternative evaluations based on different label-set modifications (as shown in the "C_{modif}" column of the Tables 5 and 6). These are i) All: Keep all the labels that appear across the EHRs from the subset of the given specialty. ii) Spec: Consists in taking into account only the "specialty labels", that is, keep the labels of the given specialty. For example, for Cardiology, you will keep only those labels that appear in Chapter IX - Diseases of the circulatory system of the ICD, due to it being the most-related chapter to Cardiology.

To help the reader interpret the results, note that while the records within Pneumology service convey 2, 057 different ICDs (see $|C_{train}|$ column from Table 5), the sub-set of ICDs within the ICD chapters related to Pneumology are 126, and those among the 110 most frequent codes are only 17 (see $|C_{test}|$ column with the 'Spec' $|C_{modif}|$ from Table 5).

First of all, in most clinical services, when evaluating with the "specialty labels", performance improves. In some specialties, such as Nephrology, the increase is considerable (26.3 F-Score points). One reason is that the difference in the number of labels between both label-set modifications is notorious in every medical service (i.e. from 107 to 17 in Pneumology, or from 105 to 23 in Cardiology…). Despite this fact, in some specialties, better performance is achieved with all the labels than only with the specialty labels, such as in the Digestive case (i.e. 32.39 F-Score with all the 108 labels, and 23.02 with only the 6 digestive-related labels). However, the most remarkable aspect is that, regarding all the labels, (the 'All' rows on the C_{modif} column from Tables 5 and 6) for all medical services, the generalist model achieves a better performance in comparison with the Specialty Models. Nonetheless, what behaviour takes place when considering only the 'specialty labels'?

It can be observed (in Tables 5 and 6) that when the aim is to classify the EHRs of a given specialty according to the ICD codes of that specialty, it is worth training the Specialty Models. In four of the six specialties, the results improve in terms of F-Score. Besides, note that for the medical services which perform better with the generalist model (Pneumology and Nephrology) the gain is only around 3 and 6 points respectively. However, the mean improvement obtained with the Specialty Models for the other specialties is about 14 points, presenting some notable increases, such as the ∼30 points improvement (from 21.28 to 52.38) in the Digestive specialty. Figure 2 combines the experiments with the generalist model evaluated per specialty and each of the Specialty Models (i.e. the results from Tables 5 and 6). The picture shows that assessing all the labels, the generalist model is more suitable, without modification, while when considering only the "specialty labels", the Specialty Models line (in light blue), is, in most medical services, superior to the generalist model.

V. DISCUSSION

The experimental setup consists of a popular clinical multi-label dataset, namely, MIMIC-III, used to validate our approximation and compare with previous works, along with some in-house datasets (Osa1 and Osa2), that presents the advantage that has the data segmented by year and medical specialty.

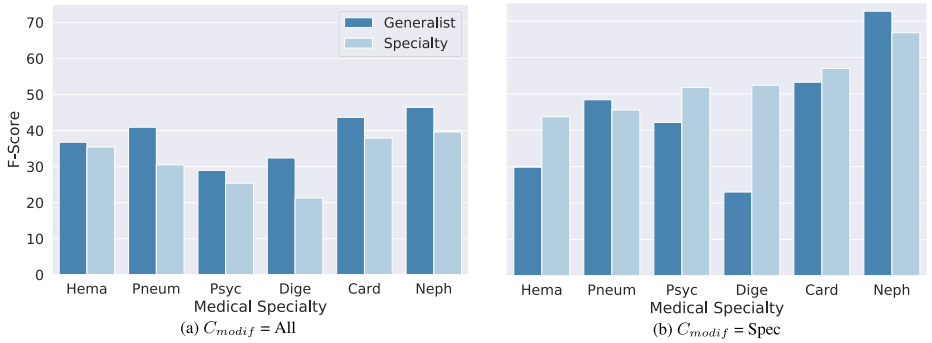


FIGURE 2. F-Scores of models trained with the Osa1+2 data-set on the $\text{Full}_{1\%}$ label-set, but re-evaluated per specialty (i.e. generalist model, dark blue) and trained on subsets of EHRs (i.e. Specialty Models, light blue).

Compared to the related works, the input EHR is not restricted to a diagnostic phrase of few words as in [49] or a short note as in [48]. Our EHRs comprise full notes with 864 ± 415 words on average it is close to a full clinical history of MIMIC-III, entailing several notes for a given patient, with $1,399 \pm 721$ words per history on average. Due to the complex ICD system in conjunction with long medical texts, nowadays, numerous healthcare professionals specially trained are devoted to manual EHR coding. Emerging techniques in NLP are bridging the gap between manual and automatic ICD coding through clinical decision support systems.

In an attempt to assess the usability of the models in practice, we should bear in mind whether the errors produced by the system are minor, due to confusion between non-essential modifiers (i.e. the last characters from the Full codes) yet correctly guessing the main class, or major, failing even the chapter of the ICD. While confusion between non-essential modifiers is counted as a failure, in practice, the system can help the human-coder position in the right branch of the ICD; by contrast, confusion between chapters would require an extra effort on the part of the human expert. In this scenario, the model would guide the expert to a Chapter (branch) of the ICD in which to select the Full code.

The assessment was carried out enabling the model dispense with the non-essential modifiers (and focusing on the main ICD and the chapter). Results showed that the potential of the model increased substantially from the fully- automated scenario to that of the computer-aided classification. A computer-aided ICD classification system can help the human expert to access the chapters of the ICD involved in each record very accurately (with a Precision of 80.88 and a coverage-recall of 67.46) as shown in Table 3. If we turn to a finer-grained classification in a situation in which the system would act automatically and would code the fully-specified label, the system would attain Precision of 27.64 and the Recall decays to 20.52. Depending on the Recall required, the system would demand a more active role from the human, while a significant percentage of the labels would have been correctly assigned. Although we have explored several levels of granularity, namely, Chapter, Main and Full granularity, we have focused on the complete ICD, as it is of great importance for applications such as insurance billing or other clinical information extraction tasks.

Often, previous works discarded learning ICDs that had little prevalence in the set or which only focused on a set of nearly a hundred labels [34], [48], [50]. In our case, we assessed them all, but as we could expect, prevalent ICDs are predicted more accurately than the average prediction quality. Table 3 disclosed that increasing training instances significantly benefitted the predictive ability of the model. The restriction to 110 and 16 most prevalent labels was selected in an attempt to make fair comparisons with previous works. Nevertheless, increasing the number of labels and decreasing the performance does not show a linear relationship, but rather the performance drop is less than expected. This implies that the model can learn from infrequent occurrences and can predict uncommon ICDs.

An **analysis** of the results shows that a generalist model trained overall services achieves, on average, an F-Score of 22.48 for the full set of 2,554 labels (see Table 3). Regarding the experiments to assess the robustness across-time, adding more years (i.e. more EHRs) to the training set benefits performance, as expected. Nevertheless, it is interesting to note that although the models are robust enough to correctly classify some EHRs from future years trained solely on data from past years, there is a negative effect on performance. Therefore, our recommendation is, whenever possible, to continue retraining the models with the new data as it becomes available since the improvement is not negligible.

In what concerns the experiments with the different medical services, one conclusion is that when evaluating the labels without modification by service (that is, all the labels that

appear in the EHRs), the best results are obtained with the generalist model, meaning that the lexical reduction did not overcome the label-set variability. Nevertheless, the most significant insight gained is derived from Table 6, which shows that it is when the Specialty Models are trained on the specialty label-sets that they do better than the generalist model. It is true that this comes with the extra cost of training several models, one for each medical service, and is limited to more restricted specialty-related label-sets. However, we feel that for certain applications, such as intra-specialty pharmacovigilance services in hospitals, these costs could be offset by the associated advantages.

Regarding the evaluation, we believe that there are aspects that do not get reflected in the most widely used metrics (such as Precision, Recall, F-Score, MAP, MRR…). Specifically, the number of codes associated with a document is relevant: therefore, if the prediction yields either a much lower, similar or much higher number of codes than the actual number of codes, this should be penalised/acknowledged accordingly. We feel that further multifaceted metrics should be developed to gain a deeper insight into extreme multi-label classification.

There is room for improvement by exploring other neural approaches e.g. models based on the Transformer architecture (BERT, BioBert, …). Nevertheless, transformers pose challenges in the training process [51] due to the high computational burden and data needed. To remedy this, and inspired by the fine-tuning strategy, we feel that an initial generalist model could be trained; this could be further fine-tuned with new data from subsequent years or new medical services.

VI. CONCLUSION

This work deals with an extreme multi-label classification task on clinical texts. The aim is to assign, to each EHR, the corresponding diagnoses as in the ICD. Each EHR tends to convey 5.8 ± 3.4 ICDs (out of about 2, 500 distinct diagnoses in our study).

Having demonstrated the ability of the approach to be both a fully-automatic and computer-aided multi-label classification, we assessed the resilience of the model to natural variations in order to address omission in previous works. The concern is about the behaviour of a model trained with some EHRs when it comes to classifying EHRs later on (e.g. texts possibly written by different experts). We put our focus on variations in two aspects: **across-time** and through **hospital services**.

Regarding the resilience of the system to time-related variations, the results showed that the datasets from different (non-overlapping and consecutive) years are similar in difficulty, as the results with Osa1 and Osa2 are reasonably comparable, with 20.43 and 18.56 F-Score, respectively. Also, adding more samples is always useful, as this gives best results, as previously seen when the train set is the union Osa1+2 (i.e. 22.48 F-Score when training and testing with Osa1+2). A key insight is that although the datasets are similarly difficult when trying to predict future EHRs with data from previous years, the performance decline significantly (i.e. from 20.43 when training and testing with Osa to 11.61 when training with Osa1 but testing with Osa2 future samples).

With respect to the ability of the system to classify EHRs across medical specialties, our experiments showed that when the predictions are made over non-modified label-sets by specialty, the best option is the generalist model, which benefits from the greater number of EHRs in the training set. However, the approach which achieves the most favourable results on specialty labels is to train Specialty Models with the specialty label sub-sets. Although this carries an extra cost, it can be useful for the development of tools for application in specific medical services in hospitals.

We feel that there is still a gap in the literature that could be exploited for **future work**: namely, knowledge-driven reinforcement learning exploiting the hierarchical structure of the ICD to gain accuracy in different granularity levels. Previous works tried to incorporate the hierarchy [52], [53]. Clinical entity recognition could help to recognise relevant information such as disorders or findings, laterality, severity or body-part. This information is, somehow, included in the hierarchical representation of the ICD and could drive the code generation. Within our framework, the hierarchical boundaries could be modelled as embedded graphs. This approach, however, is outside of the scope of this work.

REFERENCES

[1] P. Mukherjee and A. Mukherjee, "Advanced processing techniques and secure architecture for sensor networks in ubiquitous healthcare systems," in *Sensors for Health Monitoring*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 3–29.

[2] A. Banerji, K. H. Lai, Y. Li, R. R. Saff, C. A. Camargo, K. G. Blumenthal, and L. Zhou, "Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions," *J. Allergy Clin. Immunol., Pract.*, vol. 8, no. 3, pp. 1032–1038, 2020.

[3] S. Santiso, A. Perez, and A. Casillas, "Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 5, pp. 2148–2155, Sep. 2019.

[4] L. Chan, K. Beers, A. A. Yau, K. Chauhan, Á. Duffy, K. Chaudhary, N. Debnath, A. Saha, P. Pattharanitima, J. Cho, P. Kotanko, A. Federman, S. G. Coca, T. Van Vleck, and G. N. Nadkarni, "Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients," *Kidney Int.*, vol. 97, no. 2, pp. 383–392, Feb. 2020.

[5] D. Chandran, D. A. Robbins, C.-K. Chang, H. Shetty, J. Sanyal, J. Downs, M. Fok, M. Ball, R. Jackson, R. Stewart, H. Cohen, J. M. Vermeulen, F. Schirmbeck, L. de Haan, and R. Hayes, "Use of natural language processing to identify obsessive compulsive symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder," *Sci. Rep.*, vol. 9, no. 1, pp. 1–7, Dec. 2019.

[6] A. Bhattacherjee, S. Roy, S. Paul, P. Roy, N. Kausar, and N. Dey, "Classification approach for breast cancer detection using back propagation neural network: A study," in *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2020, pp. 1410–1421.

[7] N. C. Ernecoff, K. L. Wessell, L. C. Hanson, A. M. Lee, C. M. Shea, S. B. Dusetzina, M. Weinberger, and A. V. Bennett, "Electronic health record phenotypes for identifying patients with late-stage disease: A method for research and clinical application," *J. Gen. Internal Med.*, vol. 34, no. 12, pp. 2818–2823, Dec. 2019.

[8] H. Alzoubi, R. Alzubi, N. Ramzan, D. West, T. Al-Hadhrami, and M. Alazab, "A review of automatic phenotyping approaches using electronic health records," *Electronics*, vol. 8, no. 11, p. 1235, Oct. 2019.

[9] *International Statistical Classification of Diseases and Related Health Problems*, World Health Org., Geneva, Switzerland, 2004, vol. 1.
[10] K. J. O'Malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health Services Res.*, vol. 40, no. 5, pp. 1620–1639, Oct. 2005.
[11] R. H. Napier, L. S. Bruelheide, E. T. K. Demann, and R. H. Haug, "Insurance billing and coding," *Dental Clinics North Amer.*, vol. 52, no. 3, pp. 507–527, Jul. 2008.
[12] M. Lau, J. L. Prenner, A. J. Brucker, and B. L. VanderBeek, "Accuracy of billing codes used in the therapeutic care of diabetic retinopathy," *JAMA Ophthalmol.*, vol. 135, no. 7, pp. 791–794, 2017.
[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160035.
[14] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than English: Opportunities and challenges," *J. Biomed. Semantics*, vol. 9, no. 1, p. 12, Dec. 2018.
[15] H. Dalianis, *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer, 2018, doi: 10.1007/978-3-319-78503-5.
[16] D. Zikos and N. DeLellis, "CDSS-RM: A clinical decision support system reference model," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 137, Dec. 2018.
[17] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.
[18] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, and P. Zweigenbaum, "CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian," in *Proc. CLEF (Working Notes)*, 2018, pp. 1–18.
[19] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. Kamath, "Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," *Knowl.-Based Syst.*, vol. 190, Feb. 2020, Art. no. 105321.
[20] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling," 2016, *arXiv:1611.06639*. [Online]. Available: http://arxiv.org/abs/1611.06639
[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555
[22] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[23] A. Névéol, K. B. Cohen, C. Grouin, T. Hamon, T. Lavergne, L. Kelly, L. Goeuriot, G. Rey, A. Robert, X. Tannier, and P. Zweigenbaum, "Clinical information extraction at the CLEF eHealth evaluation lab 2016," in *Proc. CEUR Workshop*, vol. 1609, 2016, p. 28.
[24] A. Névéol, A. Robert, R. Anderson, K. B. Cohen, C. Grouin, T. Lavergne, G. Rey, C. Rondet, and P. Zweigenbaum, "CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French," in *Proc. CLEF (Working Notes)*, 2017, pp. 1–17.
[25] A. Dörendahl, N. Leich, B. Hummel, G. Schönfelder, and B. Grune, "Overview of the CLEF eHealth 2019 multilingual information extraction," in *Proc. CEUR-WS*, 2019, pp. 1–9.
[26] L. Zhou, C. Cheng, D. Ou, and H. Huang, "Construction of a semi-automatic ICD-10 coding system," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–12, Dec. 2020.
[27] A. Casillas, A. D. de Ilarraza, K. Gojenola, M. Oronoz, and A. Pérez, "First approaches on Spanish medical record classification using diagnostic term to class transduction," in *Proc. 10th Int. Workshop Finite State Methods Natural Lang. Process.*, 2012, pp. 60–64.
[28] A. Pérez, A. Atutxa, A. Casillas, K. Gojenola, and Á. Sellart, "Inferred joint multigram models for medical term normalization according to ICD," *Int. J. Med. Informat.*, vol. 110, pp. 111–117, Feb. 2018.
[29] O. Luaces, J. Díez, J. Barranquero, J. J. D. Coz, and A. Bahamonde, "Binary relevance efficacy for multi-label classification," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 303–313, 2012.
[30] A. Rios and R. Kavuluru, "Neural transfer learning for assigning diagnosis codes to EMRs," *Artif. Intell. Med.*, vol. 96, pp. 116–122, May 2019.
[31] C. J. Murray, A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baquim, L. Dandona, E. Dantzer, V. Das, U. Dhingra, and A. Dutta, "Population health metrics research consortium gold standard verbal autopsy validation study: Design, implementation, and development of analysis datasets," *Population Health Metrics*, vol. 9, no. 1, p. 27, Dec. 2011.
[32] W. B. A. Karaa, A. S. Ashour, D. B. Sassi, P. Roy, N. Kausar, and N. Dey, "MEDLINE text mining: An enhancement genetic algorithm based approach for document clustering," in *Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems*. Cham, Switzerland: Springer, 2016, pp. 267–287.
[33] R. Babbar and B. Schölkopf, "Data scarcity, robustness and extreme multi-label classification," *Mach. Learn.*, vol. 108, nos. 8–9, pp. 1329–1351, Sep. 2019.
[34] J. Pérez, A. Pérez, A. Casillas, and K. Gojenola, "Cardiology record multi-label classification using latent Dirichlet allocation," *Comput. Methods Programs Biomed.*, vol. 164, pp. 111–119, Oct. 2018.
[35] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5812–5822.
[36] R. You, Z. Zhang, S. Dai, and S. Zhu, "HAXMLNet: Hierarchical attention network for extreme multi-label text classification," 2019, *arXiv:1904.12578*. [Online]. Available: http://arxiv.org/abs/1904.12578
[37] F. Gargiulo, S. Silvestri, M. Ciampi, and G. De Pietro, "Deep neural network for hierarchical extreme multi-label text classification," *Appl. Soft Comput.*, vol. 79, pp. 125–138, Jun. 2019.
[38] Y. Deng, A. Sander, L. Faulstich, and K. Denecke, "Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders," *Artif. Intell. Med.*, vol. 93, pp. 29–42, Jan. 2019.
[39] A. Blanco, O. Perez-de-Viñaspre, A. Pérez, and A. Casillas, "Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity," *Comput. Methods Programs Biomed.*, vol. 188, May 2020, Art. no. 105264.
[40] Y. Cheng, K. Qian, Y. Wang, and D. Zhao, "Missing multi-label learning with non-equilibrium based on classification margin," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105924.
[41] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in EU legislation," 2019, *arXiv:1905.10892*. [Online]. Available: http://arxiv.org/abs/1905.10892
[42] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-BERT: Extreme multi-label text classification with BERT," 2019, *arXiv:1905.02331*. [Online]. Available: http://arxiv.org/abs/1905.02331
[43] Y.-T. Zhou and R. Chellappa, "Computation of optical flow using a neural network," in *Proc. IEEE Int. Conf. Neural Netw.*, Jul. 1988, pp. 71–78.
[44] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Oxford, U.K.: Butterworth-Heinemann, 1979.
[45] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, MA, USA: Cambridge Univ. Press, 2008.
[46] M. Apidianaki, S. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, "Proceedings of the 12th international workshop on semantic evaluation," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 1–18.
[47] F. Duarte, B. Martins, C. S. Pinto, and M. J. Silva, "Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text," *J. Biomed. Informat.*, vol. 80, pp. 64–77, Apr. 2018.
[48] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, and N. Taright, "Supervised topic models for diagnosis code assignment to discharge summaries," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2016, pp. 485–497.
[49] A. Atutxa, A. Casillas, N. Ezeiza, I. Goenaga, V. Fresno, K. Gojenola, R. Martinez, M. Oronoz, and O. P. D. Viñaspre, "IxaMed at CLEF eHealth 2018 task 1: ICD10 coding with a sequence-to-sequence approach," in *Proc. CLEF Online Work. Notes. CEUR-WS*, 2018, pp. 1–9.
[50] P. Nigam, "Applying deep learning to ICD-9 multi-label classification from medical records," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2016.
[51] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," 2020, *arXiv:2004.08249*. [Online]. Available: http://arxiv.org/abs/2004.08249
[52] J. C. Ferrao, F. Janela, M. D. Oliveira, and H. M. G. Martins, "Using structured EHR data and SVM to support ICD-9-CM coding," in *Proc. IEEE Int. Conf. Healthcare Informat.*, Sep. 2013, pp. 511–516.
[53] L. Cao, D. Gu, Y. Ni, and G. Xie, "Automatic ICD code assignment based on ICD's hierarchy structure for Chinese electronic medical records," *AMIA Summits Transl. Sci. Proc.*, vol. 2019, p. 417, May 2019.

ALBERTO BLANCO received the B.S. and M.S. degrees in computer engineering and computational engineering and intelligent systems from the University of The Basque Country (UPV/EHU), in 2018 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electricity and Electronics.

His research activity is being carried out within Ixa Group, HiTZ: Basque Center for Language Technology. His research interests include deep learning, natural language processing, clinical text mining, and multi-label classification.

ALICIA PÉREZ received the B.S. and M.S. degrees in physics engineering from the University of The Basque Country (UPV/EHU), and the Ph.D. degree in computational linguistics in 2010. Since 2011, she has been an Assistant Professor with the Computer Languages and Systems Department. Her research activity is being carried out within Ixa Group, HiTZ: Basque Center for Language Technology. Her research interests include natural language processing and understanding, clinical text mining, and artificial intelligence.

ARANTZA CASILLAS received the B.S. degree in computer science from Deusto University, and the Ph.D. degree in computational linguistics in 2000. Since 2001, she has been with the Science and Technology Faculty (UPV/EHU), Electricity and Electronics Department, where she is currently an Associate Professor. She is also the main Researcher of the Medical and Legal Domains Section, HiTZ: Basque Center for Language Technology. Her research interest includes artificial intelligence, natural language processing and understanding, clinical text mining, and clinical information retrieval.

• • •

1374 IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 26, NO. 3, MARCH 2022

Exploiting ICD Hierarchy for Classification of EHRs in Spanish Through Multi-Task Transformers

Alberto Blanco, Alicia Pérez, and Arantza Casillas

***Abstract*—Electronic Health Records (EHRs) convey valuable information. Experts in clinical documentation read the report, understand the prior work, procedures, tests carried out, and encode the EHRs according to the International Classification of Diseases (ICD). Assigning these codes to the EHRs helps to share information, and extract statistics. In this paper, we explore computer-aided multi-label classification approaches. While Natural Language Understanding has evolved for clinical text mining, there is still a gap for languages other than English. Language-modeling aware Transformers has demonstrated state of the art approaches through exploiting contextual dependencies. Here we focus on EHRs written in Spanish, and try to benefit from the Language Model itself, with unannotated corpus with less data but in-house, in-domain and closely-related EHRs to that of the downstream task. The International Classification of Diseases coding scheme is hierarchical, but its synergies among hierarchical levels are rarely exploited. In this work, we implement and release a hierarchical head for multi-label classification, which benefits from the hierarchy of the ICD via multi-task classification.**

***Index Terms*—Hierarchical, transformers, multi-task, EHR, multi-label, under-resourced languages.**

I. Introduction

THIS work deals with the coding of Electronic Health Records (EHRs) according to the International Classification of Diseases (ICD). This task requires, on the one hand, a well-suited Language Model (LM) to cope with the contextual nuances of the language and, on the other hand, an extreme multi-label classifier to assign each EHR a subset of ICD codes from the thousands of diseases available in the system. In this task, we focus on the classification of EHRs in Spanish. In recent years, there has been an international interest in this area, and since 2012 the CLEF eHealth Evaluation Lab and Workshop Series have been organized every year, with many ICD coding tasks, focusing mainly in English. Regarding the minority languages, there have been editions focused on French, Hungarian and Italian [1], [2]. Concerning Spanish, only the 2020 edition [3], task 1 [4] focused on clinical text term coding from clinical case records and the CIE-10-ES, the Spanish version of the ICD-10.

The *Transformer* architecture [5] was conceived to tackle the sequence transduction problem, i.e., any task that transforms an input sequence, the text from the EHR, to an output sequence, an array. The LM component is in charge of transforming the textual input tokens into numeric vectors, which is the optimal form to compute the probabilities of each ICD in the classification module. Transformers follow an encoder-decoder pattern, in which each encoder component is built of a self-attention and a feed-forward neural network module. The decoder structure is similar, but with an added attention module. An encoder-decoder pair is known as Transformer. The objective of self-attention mechanism is to relate different positions of a single input sequence, and have proven to extend the ability of RNNs to model dependences to long-distance patterns.

Pretraining LMs on large amounts of general domain unlabeled text [6], [7] have had a big impact on the performance in a variety of Natural Language Processing (NLP) tasks. In recent years LMs as ELMo [8] and, more recently, BERT [9] have demonstrated the highest competence in Natural Language Understanding (NLU).

Apart from the LM component itself, the model also portrays a neural-network-based module for the classification. The LM takes care of the extraction of the contextual features that are later passed onto the classification module. For that reason, the quality of the embeddings representations generated by the LM are vital. Nevertheless, these models, assembled as Deep Neural Networks still depend on significant quantities of annotated data [10] for tasks as text classification [11], [12]. The pre-training of LMs in general and, more specifically, BERT models, from both general and in-domain data are used to improve further downstream tasks; which is is an open NLU question. [13] investigated how the pre-trained BERT can be adapted for biomedical corpora, and introduced BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), a domain-specific LM pre-trained on large-scale biomedical corpora. When it comes to specialized domains like biomedicine, incorporating in-domain text enhances the performance of the system as the generalist LM is adapted to

Manuscript received January 8, 2021; revised June 17, 2021 and July 28, 2021; accepted July 29, 2021. Date of publication September 14, 2021; date of current version March 7, 2022. This work was supported in part by the Spanish Ministry of Science and Technology (PAT-MED PID2019-106942RB-C31) and in part by the Basque Government (Elkartek KK-2019/00045, IXA IT-1343-19, Predoctoral under Grant PRE-2019-1-0158). *(Corresponding author: Alberto Blanco.)*

The authors are with the HiTZ Center - Ixa, University of the Basque Country UPV/EHU, 20080 Donostia, Spain (e-mail: alberto.blanco@ehu.eus; alicia.perez@ehu.eus; arantza.casillas@ehu.eus).

Digital Object Identifier 10.1109/JBHI.2021.3112130

the domain [13]–[16]. In our work, we take the BioBERT model and the generalist BERT Multilingual model as base models. We opted for BERT as it outperformed other models in the related BLUE benchmark [15].

The transfer learning strategies such as the pre-training of LMs are of particular importance to minority languages and on fields with high lexical variability. The case of EHR classification in Spanish suffers from both, as it is not a broad studied language in NLP, and there are insufficient clinical notes publicly available [3].

The ICD code system arranges, hierarchically, thousands of codes from general (or coarse-grained) information, to fully-specified codes incorporating non-essential modifiers. The main task of this work is the classification of EHRs according to the ICD codes that are present on a given medical record. Discriminating between thousands of codes is challenging for automated systems. Particularly, the consistency between predictions remains as an open research question in the field. Indeed, the most extended binary relevance approach has been criticized, precisely, for disregarding relationships among labels [17]. However, the multi-label classification head applied in this work handles the relationships among labels within the same hierarchy level, as it computes all the probabilities with the Sigmoid layer. Moreover, we exploit the hierarchy to improve the ICD coding systems and, specifically, to further improve the consistency between predictions, benefiting from relationships between a label and its descendants at different levels of the hierarchy.

To that end, the hierarchy is exploited through a multi-task BERT-based transfer learning approach. Hence, instead of applying a single-task (regular) classification architecture, we propose a multi-task classification module. The difference is that the model handles several classification tasks simultaneously, i.e., the model has three independent classification heads trained together but with shared parameters that allow synergies between the tasks. The three classifications sub-tasks must be related, share computed sub-features and be reasonably dependent [18], [19]. The insights that the model learns from each sub-task are distinct and experimental results show that mixing them into the same model lead to improved overall results.

Each sub-task requires a unique label-set related to each EHR to leverage the multi-task learning paradigm, but obtaining thousands of EHRs labelled is a painstaking and expensive task. We propose to exploit the hierarchical nature of the ICD to extract the required label-sets to allow multi-task learning. Precisely, from each full ICD label, we extract two more labels by cutting the label based on some hierarchy levels of the ICD (i.e., from the full "I25.110" ICD code, we would also extract the "I25" label and the "I" label). We refer to these three levels as Full, Main, and Chapter label levels, as presented in Figure 1 in Section III.

That way, we promote cooperation between different levels from a given branch of the ICD and competence between different branches. For example, if the system predicts a diagnosis from Chapter IX of the ICD code (Diseases of the Circulatory System), a more specific diagnosis code related to circulatory diseases could be predicted. However, if the model predicts that there are no diagnosis related with the Chapter IX, a more

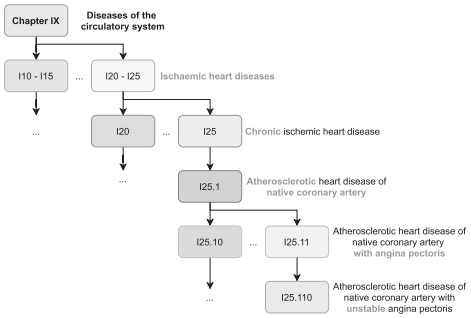


Fig. 1. Example of hierarchy in the structure of ICD codes, taking the I25.110 code and its predecessors in the ICD hierarchy as example.

specific code related with Diseases of the Circulatory System should not be predicted. The multi-task learning paradigm shares relevant information across related tasks to the benefit of each of the single tasks involved [20], [21]. Multi-task learning is particularly convenient when the annotated available data is scarce [22], as is the usual case in the biomedical field and even more so in languages other than English [23], [24]. Following these ideas, we propose a multi-label multi-task text classifier with a pre-trained BERT-based core with an enhanced LM with biomedical text data, for the ICD classification of EHRs.

Moreover, this applies to every multi-label set which could be interpreted as a hierarchy. In a hierarchical label-set, each label can be unfolded in several labels of different granularity, i.e., the level of detail of the label, or in other words, the depth of the tree that describes the given hierarchical label-set. The **motivation** of this work is twofold: 1) Leverage the vast amounts of unannotated (i.e., unsupervised, with no coded labels) biomedical data to adapt the LM. 2) Exploit the hierarchical nature of the ICD to improve label confidence. In short, we hypothesize that even without further annotated data, that could help to improve the classification module, the system can be benefited from small though specific unannotated corpora to improve the underlying LM to the benefit of the classification module. Second, from the hierarchical structure of the ICD insights can be gained to improve the confidence of the extreme multi-label classification system without virtually incurring in more computational costs.

II. Related Work

Automatizing EHR codification would be desirable, nevertheless, it remains a challenging task. The prior work faced the task in original, though hardly comparable or reproducible ways: with alternative assessment metrics, the number of EHRs and size of the label set and with English as the dominating language. Regarding ICD multi-label classification, [25], explored a combination of example-based methods to capture codes with varying prevalence and employed representations based on semantic and lexical features. This work involved 7,078

ICDs and only 5,803 training EHRs [25] while [26] involved 5,324 ICDs and 26,373 training EHRs. Our dataset is closer to the latter but it is in Spanish like the former.

Extreme multi-label classification problem was tackled by virtue of millions of in-house EHRs coded with ICD-10 [27]. This approach applied a BERT model trained from scratch on EHR notes and adapted the BERT architecture with a multi-label attention system.

The vital issue in our work is to exploit the hierarchical nature of the ICD in order to enhance the learning procedure and the LM inferred from the EHRs. The hierarchical structure of ICD has also been studied in the previous work either with classical methods as Hierarchical Support Vector Machines [28] or Graph Neural Networks [29] for ICD coding. However, we aim to exploit the hierarchy with the current state-of-the-art Deep Learning models, i.e., via the Transformers architecture. [26] addresses the EHR multi-label classification according to the ICD-10 code considering 23,000 6-digit codes and about 1,900 corresponding 3-digit codes (respectively, similar to our Full and Main labels). To take advantages of the hierarchy of the ICD, they proposed a two-stage framework which first predicts the Main category codes, and then searches the specific or Full category. A drawback is that their subcategory models do not consider all the category codes, only subsets instead, this makes the model unaware of relationships among labels. On the other hand, hierarchy was also exploited by stacking new sets of classifiers at each level [30]. While this approach is intelligent, requires low cardinality label-sets, such as the 9 labels reported by the authors, but it becomes intractable when the number of labels is high, as it is our, especially with memory and computationally intensive models as the Transformers [31]. Multi-tasking offered us an alternative to address this issue.

The Structured Output Learning methods are an alternative way of tackling multi-label learning in a non-flat fashion, as they rely on an output graph connecting multiple labels to model the correlation between labels [32]. In that sense, it follows a similar idea as our hierarchical multi-task setting – Hierarchical classification is a specific approach of Structured Output Prediction. Similarly, we apply a multi-task classification model that predicts three hierarchy levels in parallel, with the predictions of lower levels influencing the higher levels of the hierarchy. However, there are significant methodological differences; for example, the Structured Support Vector Machines (SSVM) perform the learning by using discriminant function over input-output pairs [33]. That way, SSVMs can predict complex objects like trees, sequences or sets, but our multi-task method relies on flat multi-label heads trained concurrently, whose predictions are later combined, following the hierarchy. Moreover, our architecture allows having both shared parameters (in the LM part of the architecture) and specific parameters for each sub-task (in the classification heads).

The effect of pre-training and fine-tuning LMs with in-domain data was studied in the form of both domain adaptation, and task adaptation by the prior work [27], [34], [35]. So far, however, there has been little discussion about enhancing the LM with small though specific unannotated corpora. To that end we bet on a three-step strategy where the novel step consists of further fine-tuning the LM with unannotated data but with the source of this data being the same as in the training of the downstream task. Furthermore, to overcome the computational problems that can arise with the approaches from [26], [30], but while also benefiting from the hierarchical structure of the ICD code, we proposed the multi-task strategy. First, the computation of the probabilities for each label-set are obtained in parallel. Second, the relationships from all the labels are preserved. That is, the Chapter predictions influence the Main predictions, and the Main predictions influence the Full predictions, but at each label-set, all the labels are considered. With this, the new sets of labels required from the multi-task are extracted respecting the hierarchical levels.

The comparison with previous works is challenging due to the disparity and privacy of most datasets and the lack of standard datasets and benchmarks [34]. It is relevant to consider aspects as the language or the type of electronic health record, but even when applying the same dataset, there are substantial variations due to preprocessing that lead to changes in the labels' minimum-frequency or the assessment metric utilised. Some works classify based on the ICD-9 [26], while others apply the updated ICD-10 version, as in [4], [25], [27] and on this work, and some works apply modifications of the ICD, such as the ICD-O (Oncology variation) as [30]. This work employs EHRs in Spanish, as [4], [25], but others like [15], [26], [27], [30] work with English EHRs. Regarding the cardinality, there are substantial differences, from the 9 labels of [30] to the 7,078 of [25]; our largest dataset has 3,656 ICD codes. Also, among the works that use Transformers and pre-train the models, there are differences in the strategies. [27] pre-train only with EHRs, while [15] apply EHRs plus biomedical articles, and we apply a general domain pre-training in addition to EHRs from a similar dataset to the ones used for our classification task. The previous works that also introduced the hierarchy in their architectures got 91.16 [26] and 91.8 [30] in terms of F1-Score micro.

III. Materials

An ICD-10 label is made of 3 to 7 alpha-numeric characters, and it is arranged hierarchically in branches. Starting from the lowest level of granularity, we find what is called Chapter (i.e., Chapter IX, referred to with the character "I," following the example from Figure 1). The Chapter is encoded in this structure as the first character and reflects a coarse-grained classification of the EHR that roughly translates into medical specialities [36]. Besides, we can take the first three characters (i.e., I25), considered the Main ICD class, (from now on just Main in short). The remaining characters comprise of what is referred to as non-essential modifiers (e.g. laterality and severity, i.e., the last "110" from I25.110), which are referred to as the "fully-specified ICD class," we will refer to it as Full.

The materials used can be classified as the unannotated dataset applied to fine tune the LM and the annotated dataset to train the multi-label classifier.

The **unannotated dataset** is a collection of clinical documents extracted from the same source (a hospital) as the supervised dataset. Therefore, both datasets are comprised of EHRs in

TABLE I
QUANTITATIVE DESCRIPTION OF THE ANNOTATED DATASET, CONSIDERING THREE HIERARCHICAL LEVELS (H-LEVELS) FROM THE ICD (FULL, MAIN, CHAPTER). TWO PERSPECTIVES OF THE DATASET WERE CONSIDERED REMOVING THE THE ICDS BELOW A MINIMUM FREQUENCY THRESHOLD, SET AS 1% IN OSA-1R AND 0% (NO REDUCTION) IN OSA-0R

		h-level		
		Full	Main	Chap
$\mathcal{Y}$	Num ICDs	3,656	1,003	24
	Avg. Card.	5.8 ± 3.3	5.6	3.8
$\mathcal{X}$	EHRs	26,969		
	Vocab	379,121		
	Words/EHR	865 ± 415		
$\mathcal{S}$	EHRs per ICD	43 ± 217		
	ICDs per EHR	5.8		

(a) **Osa-0r** dataset

		h-level		
		Full	Main	Chap
$\mathcal{Y}$	Num ICDs	104	72	16
	Avg. Card.	3.9	3.8	2.8
$\mathcal{X}$	EHRs	22,609		
	Vocab	342,738		
	Words/EHR	890 ± 415		
$\mathcal{S}$	EHRs per ICD	841 ± 985		
	ICDs per EHR	3.8 ± 2.3		

(b) **Osa-1r** dataset

Spanish and are very similar in terms of vocabulary and syntax. This unannotated dataset comprises 194,162 records, leading to a vocabulary (number of distinct words) of 1,057,787 words. Note that comparing with prior work our unannotated dataset is small, as for example the BioBERT pre-train corpus counted with 18B words while our unannotated dataset has only $\sim$100 M words.

Table I shows a summary of the **supervised dataset**, in terms of samples ($\mathcal{S}$), input ($\mathcal{X}$), and output ($\mathcal{Y}$). To generate a more extensive experimental setup and to maintain concordance with previous works, we have applied a label-set reduction based on the minimum number of appearances of labels across the documents, following the criteria from related works [37], [38]. The 1% reduction means that we have only preserved those labels that appear at least in 1% of the clinical documents.

Regarding the number of instances, the larger dataset partition amounts to 26,969 documents, with no label-set reduction, which decreases to 22,609 when the 1% reduction is applied. These 26,969 documents lead to a vocabulary (number of unique words) of 379,121 words, which decreases along with the number of documents. Something to take into account is that although the unannotated dataset has a huge vocabulary of 1,057,787 words, and the supervised dataset just a 35.8% of that number, the amount of Out-of-vocabulary words is still considerably large: 195,536 words, which supposes a 51.6% of the entire vocabulary. Concerning the labels, there are total of 3,656 labels when considering the Full labels with no reduction, 1,003 Main labels, and 24 Chapter labels. The average cardinality (i.e., mean number of labels per document) is around 5 when no reduction is applied; dropping to 3 when considering only the 1% most frequent labels. We hypothesize that by taking advantage of the hierarchical structure of the ICD via the label granularity, we could mitigate the problems related to scarcity and imbalance in the labels, and to this end, we approached it with the multi-tasking paradigm. The datasets applied in related works, apart from the MIMIC and PubMed, are not publicly available as they comprise private and sensible data. The train and test partition was carried out following a random iterative stratification strategy, respectively, with proportions 70%, 30%.

IV. METHODS

The Bidirectional Encoder Representations from Transformers, i.e. BERT [9], is suitable for this work mainly because of the following two abilities: i) it was designed to infer LMs as bidirectional representations from unannotated text, hence, it conveys information from both left and right context, and ii) the resulting pre-trained LM can be adapted (also known as fine-tuned) for multi-label classification just using one additional output layer (also known as "head"). These two modules (pre-trained LM and fine-tuned head model) are shown in Figure 2(a). Most of the parameters inferred from the corpus lie in the LM module. For example, the ICD multi-label classification head developed for this work accounts for less than the 1% of the total model parameters (even using the smallest version of BERT, which counts with 110 M of parameters). For this work, we turned to the $\text{BERT}_{\textbf{BASE}}$ model.

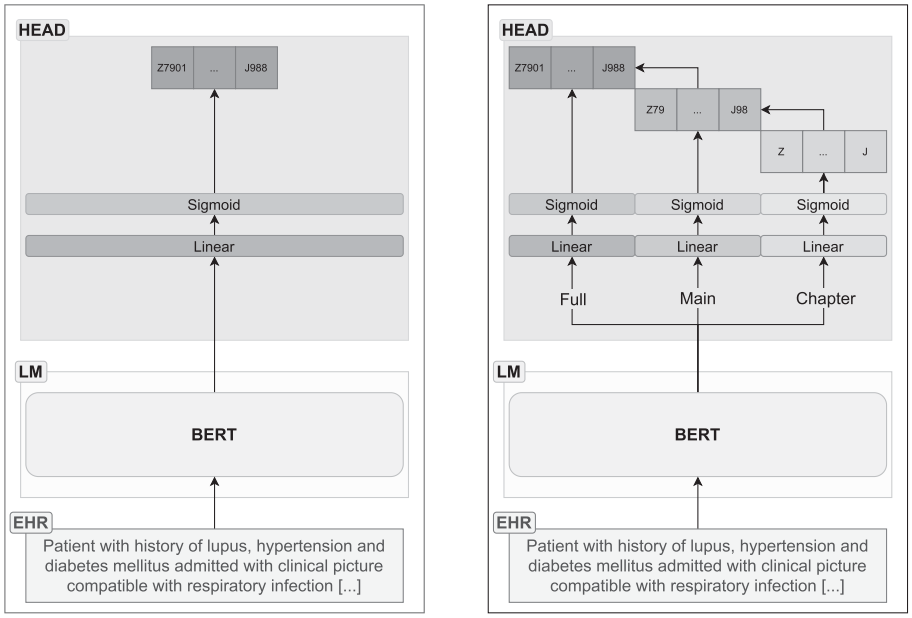
In the oncoming sections, we delve into each of the two contributions proposed to enhance multi-label classification of EHRs in Spanish.

A. Few Though Specific Data to Enhance LM

The LM module can be trained on its own (just pre-training the LM) and also in conjunction with the Head, when fine-tuning on the downstream task. The training of the LM simply requires great amounts of unannotated corpora, while the multi-label classification head can only be trained with annotated samples. Some related works employed BERT and focused on the LM's enrichment by feeding it with millions of documents, although not precisely from the same biomedical domain (e.g. abstracts from scientific articles instead of EHRs). In languages other than English, as it is our case, seldom can we find alternative sources of massive data [23]. If finding sources of unannotated data is not easy in these contexts, annotated data is essentially an illusion. Hence, while prior work explored the aid of big unannotated and annotated data to enhance, respectively, the LM and the Head, this is not a feasible approach in languages with fewer resources available. Our **hypothesis** is that even limited, though specific, unannotated data can help to further enrich the LM and to get an improved representation of the input that could significantly benefit the downstream task i.e. the multi-label classification.

B. Hierarchical Approach to Enhance Classification Module

As expert coders read the EHR they receive more and more precise information about the fully specified diagnostic term (following the example in Figure 1): a few lines from the EHR would help to restrict the diagnoses to a Chapter (e.g. "disease of the circulatory system") while further reading of details and



(a) Non-hierarchical model (b) Hierarchical model

Fig. 2. Multi-label text classification models: (a) the non-hierarchical approach was implemented as a single task architecture; (b) the hierarchical approach was implemented as a multi-task architecture exploiting ICD Chapter and Main class outcomes to get the fully-specified code to gain insights from different levels of granularity in parallel.

tests would yield the information to obtain the fully-specified diagnostic term (e.g. I25.110). Inspired by human experts we wondered if a system would also easily map data into the high levels (coarse-grained ICD labels) of the ICD hierarchy. If the model would be able to make good multi-label predictions at high-levels, then, predictions at lower levels could be driven by constraints inherited from higher-levels. The motivation behind the use of the hierarchical approach is to exploit the relationships at the ICD level in order to attain improved label confidence at fully-specified ICD level. The idea is to render the label confidence from coarse-grained to fine-grained levels in the hierarchy assuming that the model gets better predictive ability on coarse grained ICD codes.

Several works [39], [40] approach the multi-task learning paradigm to improve results leveraging the information shared among related tasks. The drawback of the approach is the fact that each singular task involved requires supervised data, which usually is difficult and expensive to obtain. For example, [39] proposed bidirectional LMs based on the multi-task learning method for text classification to prevent the shared and private spaces (in our work, the LM and multi-label head components correspondingly) of each task from merging information from each other. They added language modelling as an auxiliary task to the private part, to improve the extraction of task-specific features while promoting the shared part to learn common features. We aim to transfer this knowledge from the general NLP area to the biomedical NLP with each task being specialized in language relevant to different levels of the hierarchy. Moreover, our method does not require additional supervised data for each task, due to the fact that the label-sets are, simply, simpler representations from upper levels of the ICD hierarchy. Likewise, there are plenty of works [40], [41] which leverage the multi-task learning for exploiting synergies among tasks, model relationships between labels and improve the results, but to best of our knowledge, at the expense of more supervised data, which is an important limiting factor that we want to overcome in this work.

Our **contribution** to boost the ICD's hierarchical nature is approached by means of a multi-task classification as shown in Figure 2(b). This required us to implement the hierarchical head and, to this end, we rely on the Huggingface Transformers library [42] and have developed a multi-label sequence classification head for BERT; which is also compatible with Roberta [6], XLNet [43], XLM and DistilBERT [44]. The implementation of the hierarchical multi-label classification head is released with

this article (see Section V-C) together with the non-hierarchical approach in an attempt to promote reproducible research.

Our head for multi-label classification consists of a linear layer taking the final hidden state ($\mathbf{h}$) followed by a dropout layer and a Sigmoid activation function, as in (1). The y_i component accounts for the estimated probability of the ICD class C_i to be present in the given input (EHR tokens), with W and $\mathbf{b}$ being the parameters of the linear layer (i.e. the downstream task-specific layer).

$$\mathbf{y} = \sigma(W\mathbf{h} + \mathbf{b}) \tag{1}$$

The fine-tuning of the model is done jointly with the LM part with an appropriate loss function for the task. In this case, Binary Cross-Entropy (BCE) loss, described as (2), was employed, with $\mathbf{x}$ being the output of the Linear and Sigmoid layers, and $\mathbf{y}$ the vector representing the presence or absence of ICD codes for the EHR given.

$$BCE(\mathbf{x}, \mathbf{y}) = -W[\mathbf{y}\log(\mathbf{x}) + (1-\mathbf{y})\log(1-\mathbf{x})] \tag{2}$$

To bring together the BERT model, the multi-label classification and the multi-task learning, we have also developed a variation of the single-task head, which allows the joint learning of multiple set of labels. The multi-task head is a composition of the single-task head, one for each set of labels. Fig. 2 shows a representation of the model architectures, non-hierarchical (Figure 2(a)) and hierarchical (Figure 2(b)). The benefit of this hierarchical multi-task architecture is that on the shared parts (the LM component) of the model is where the modelling of the label-set relationships occurs, but still treats the nuances of each label-set individually in the specific-task layers (the multi-label head). Note that, additionally, there is a label relationship modelling in cascade, as the predictions for the Chapter set of labels influence the predictions for the Main set of labels, which for their part influences the predictions for the Full labels.

For each Chapter label Ch_i guessed as present in a given EHR, the system restricts the predictions to the descendants of Ch_i in the ICD hierarchy. I.e., the model should not predict "Chronic ischemic heart disease" (i.e. ICD Main class I25) unless at the Chapter level the prediction was "Diseases of the Circulatory System" (i.e. ICD Chapter I). This correction is implemented with a mask implementing a logical AND operation between the bit stating presence or absence of the j-th Main code candidate, $y(M_j)$, and the presence-bit of its ascendant Chapter code $y(C_i)$, with $ascendant(M_j) = C_i$. The mask operation shown in (3) is a simple means of leveraging label-confidence consistent with the insights gained from each level in the multi-task approach.

$$y(M_j) \leftarrow y(M_j) \ \wedge \ y(ascendant(M_j)) \tag{3}$$

$$y(M_j) \leftarrow y(M_j) \ \cdot \ y(ascendant(M_j)) \tag{4}$$

Moreover, as an alternative to (3), we have tried a fuzzy masking strategy, taking the estimated probability at each level for weighting the estimated likelihood of descendants as in (4). I.e., instead of utilising the predictions after applying the threshold (i.e., marking the label as present or non-present), we use the raw probabilities. That is, the task devote to estimate the confidence of each Chapter code ($y(C_i)$), would re-adjust the confidence estimated by the task devoted to model Main code (each $y(M_j)$) of its descendants (i.e. with $C_i = ascendant(M_j)$). The same applies to the relations between the Main and Full labels.

Regarding the LM pre-training strategies, for this work, we obtained a general domain pre-trained $\mathrm{BERT}_{\mathbf{BASE}}$ Multilingual model (specifically the Wikipedia, as the model have been trained with 104 languages, including Spanish and English, with the largest Wikipedia dumps,[1] which we refer to it as Multi-Wiki)), and a continuous pre-trained $\mathrm{BERT}_{\mathbf{BASE}}$ model (BioBERT). From there, we further enriched both models with in-house, in-domain and closely-related clinical text, the unannotated dataset of EHRs in Spanish introduced in Section III. The enriching of the model is done through the standard pre-training procedure, i.e., we continue pre-training the model from the last checkpoint but using our own data. We hypothesise that no matter which pre-trained model is applied for the downstream task (i.e., a generalist pre-trained model, or a continuous pre-trained one), if there is a considerable amount of unannotated text available, it can be used to further enrich any already pre-trained BERT model. For example, if there is a dataset with a few thousands of supervised records from a hospital, it is probably hard to get more instances. However, it is generally easier to get hundreds of thousands of unlabelled samples with no more significant supervision effort and use it to improve the supervised task results.

V. Experimental Results

A. Results Attained Enhancing the LM

The first set of experimental results, shown in, Table II, focuses on the **non-hierarchical single-task** paradigm and delves into the hypothesis stated in Section IV-A. While LM is usually trained with Big Data, our first aim was to assess the influence of enhancing LMs with few though specific data. Thus, "LM-enriched" in Table II is set to "Yes" whenever the LM model was further pre-trained with our small set of unannotated data. The table reveals the influence of the LM pre-training on two the different BERT models. On one hand, an instance of $\mathrm{BERT}_{\mathbf{BASE}}$ Multilingual [9] with no more specific pre-training (Multi-Wiki) and, on the other hand, an instance of BioBERT, which is a model grounded on $\mathrm{BERT}_{\mathbf{BASE}}$ but further pre-trained on biomedical corpus (BioBERT [13]). Then, we further pre-trained (i.e., enriched the LM) both models on our in-house unannotated dataset (similarly to which the authors of BioBERT did, with fewer data but more related to our downstream task). Table II also reveals the performance (in terms of precision (P), recall (R) and F1-Score (F), weighted averaged) of each non-hierarchical approach (shown in Figure 2(a)) at each level of the hierarchy (Chapter, Main, Full). Note that, in this non-hierarchical approach an independent model was trained for each hierarchical level. The experiment was carried out twice, respectively for Osa-0r and Osa-1r corpora shown in Tables III a and III b. With this, the aim is to show the performance of the system on extreme multi-label classification

[1][Online]. Available: https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages

TABLE II
RESULTS, IN TERMS OF PRECISION (P), RECALL (R) AND F1-SCORE (F), SHOW THE INFLUENCE OF ENRICHING THE LM WITH FEW IN-HOUSE THOUGH CLOSELY-RELATED UNANNOTATED DATA IN THE NON-HIERARCHICAL APPROACH. THE RESULTS SHOW THE PERFORMANCE OVER THE THREE HIERARCHICAL LEVELS (FULL, MAIN AND CHAPTER)

(a) Performance of the non-hierarchical approach on **Osa-0r** corpus.

h-level	BERT base	LM-enriched	P	R	F
Full	Multi-Wiki	No	12.53	15.38	12.35
		Yes	14.78	18.13	15.6
	BioBERT	No	11.32	15.86	12.38
		Yes	**16.3**	**20.01**	**16.95**
Main	Multi-Wiki	No	32.29	36.13	33.15
		Yes	35.08	37.04	34.68
	BioBERT	No	31.64	35.76	32.64
		Yes	**36.23**	**37.32**	**35.62**
Chap	Multi-Wiki	No	71.17	68.96	69.89
		Yes	71.14	**70.06**	**70.41**
	BioBERT	No	70.85	68.92	69.62
		Yes	**71.45**	68.69	69.8

(b) Performance of the non-hierarchical approach on **Osa-1r** corpus.

h-level	BERT base	LM-enriched	P	R	F
Full	Multi-Wiki	No	48.23	45.81	46.36
		Yes	**49.29**	**47.13**	**47.62**
	BioBERT	No	44.95	45.19	44.36
		Yes	48.94	46.95	47.35
Main	Multi-Wiki	No	55.97	54.6	54.88
		Yes	**57.22**	**56.36**	**56.36**
	BioBERT	No	55.11	53.23	53.67
		Yes	57.03	55.15	55.63
Chap	Multi-Wiki	No	74.25	74.29	74.06
		Yes	73.85	**75.62**	**74.6**
	BioBERT	No	73.84	72.94	73.16
		Yes	**74.69**	73.56	73.89

(with Osa-0r) one-shot learning and also on a easier task with a smaller label-set ensuring a slightly higher repetition ratio for each label (each ICD appearing at least in the 1% of the EHRs at Osa-1r).

The main issue of this first set of experiments is to prove the value of enhancing the pre-trained LM with small though closely-related datasets (as stated in Section IV-A) and, in those terms, we can identify the first **finding**. We can observe that the LMs that were further enriched with our in-house and closely-related unannotated dataset consistently outperformed those with no further pre-training. This is the case for both the general-purpose $\text{BERT}_{\mathbf{BASE}}$ Multi [9] and the biomedical model BioBERT [13].

Logically, we can see how the results systematically improve for all the models according to the label-set size, from the Osa-0r version with Full labels (3,656 labels) to the Chapter with 1% reduction (16 labels). In the former case, the best results were attained by the BioBERT further pre-trained with the unannotated Osa dataset, but with 16.95 F-Score points, while for the latter, an F-Score of 74.6 was obtained by the $\text{BERT}_{\mathbf{BASE}}$ Multi, also further pre-trained with the Osa data.

Regarding the differences between $\text{BERT}_{\mathbf{BASE}}$ Multi and BioBERT, it seems that the general-purpose model is superior when it comes to reduced label-sets; while the biomedical model brings better results when considering the full label-sets, although the differences are marginal.

TABLE III
RESULTS ATTAINED EXPLOITING HIERARCHICAL NATURE APPROACHED AS A MULTI-TASK EXPERIMENT, WHICH TEST THE BEST PERFORMING MODELS ON THE MULTI-TASK SETTING. BOTH MODELS ARE THE LM-ENRICHED, AND THE FUZZY MASKING IS APPLIED. THE TWO LABEL-SET REDUCTIONS ARE APPLIED, NO REDUCTION AT ALL AND KEEPING ONLY THE LABELS THAT APPEAR IN AT LEAST 1% OF THE DOCUMENTS

(a) Performance of the hierarchical approach on **Osa-0r** corpus.

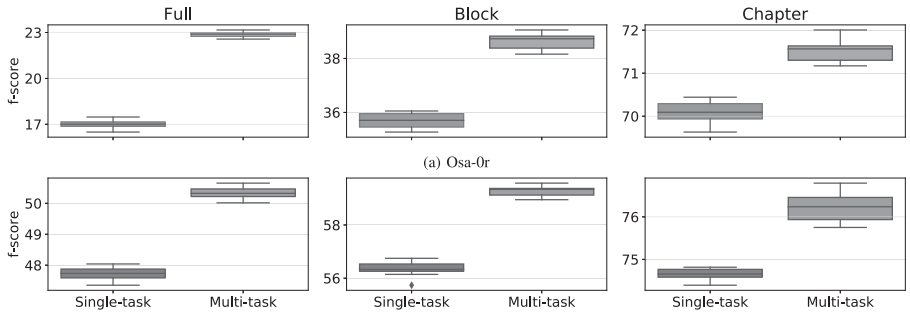
h-level	BERT base	LM-enriched	P	R	F
Full	Multi-Wiki	Yes	**21.98**	**28.36**	**22.74**
	BioBERT	Yes	13.98	34.62	17.90
Main	Multi-Wiki	Yes	**33.74**	**62.11**	**46.96**
	BioBERT	Yes	20.94	50.70	28.22
Chap	Multi-Wiki	Yes	**72.63**	**71.11**	**71.45**
	BioBERT	Yes	71.67	67.83	69.39

(b) Performance of the hierarchical approach on **Osa-1r** corpus.

h-level	BERT base	LM-enriched	P	R	F
Full	Multi-Wiki	Yes	**51.60**	**53.53**	**50.82**
	BioBERT	Yes	49.78	51.48	48.67
Main	Multi-Wiki	Yes	**56.29**	**64.16**	**59.4**
	BioBERT	Yes	54.45	61.61	57.21
Chap	Multi-Wiki	Yes	**75.83**	**76.98**	**76.21**
	BioBERT	Yes	75.07	75.38	75.06

B. Results Boosting the Hierarchy of the ICD

It is important to note that as the highest level of the hierarchy labels (Chapter) brings better results (as there are less and less specific labels), we hypothesise that the prediction of these labels in a multi-task setting can aid with the prediction of the labels from lower levels of the hierarchy, i.e., Main and, ideally, Full labels. This intuition lead us to propose the hierarchical approach (developed in Section IV-B). Moreover, before presenting the results attained with the hierarchical approach, note that in the non-hierarchical approach each level of the hierarchy involved the training of a specific system (i.e. three systems one for each of the Full, Main, Chap). By contrast, the hierarchical approach, being implemented as a multi-task model, is able to provide the estimated probabilities for each of the three h-levels and, even more, to combine these probabilities to enhance the Main and Full levels, by means of the fuzzy masking shown in (4). In brief, given that the head of the system is marginally complex in comparison to the LM, the computational complexity of the hierarchical and non-hierarchical approaches are of the same order of magnitude. More specifically, as the multi-label classification head accounts for less than 1% of the total parameters, when considering the multi-task setting, i.e., applying three analogous heads, the computational impact is marginal. The training procedure with the hierarchical model takes solely 2% more time (and memory) compared to the non-hierarchical model. Having proven that enhancing the LM brings further advantages, we explored the performance of the multi-task approach (depicted in Figure 2(b)) in Table III. For readability, we skipped results attained without enhancing the LM (i.e. LM-enriched = No) since it does nothing but corroborate aforementioned finding. Nevertheless, since there was not a clear BERT base winner from the previous experiment (in Table II), we kept showing the two BERT base models ($\text{BERT}_{\mathbf{BASE}}$ Multi and BioBERT), and fine-tuned them for the downstream task of replacing the



(a) Osa-0r

(b) Osa-1r

Fig. 3. F-score (y axis) attained by both the non-hierarchical model approached as single-task and the hierarchical model implemented as multi-task (x axis) summarising 5 runs of the experiments described in Tables II and III for Multi-Wiki with LM-enriched model.

single-task head with the multi-task one. Unlike in the previous set of experiments, here the dominance of the Multi-Wiki is clear, as it surpasses the BioBERT model in every setup and metric.

The key issue addressed in Section IV-B was to assess the aid of the hierarchical approach (see the results in Table III). Is it possible to boost the results leveraging the hierarchical fashion of the ICD by approaching the classification through a multi-task setting that has no additional supervision efforts? Figure 3 summarizes, graphically, the variations, in terms of F-score, between the non-hierarchical approach (shown in Table II) and the hierarchical one (shown in Table III). The F-scores from the hierarchical approach overcomes the non-hierarchical one in all levels of the hierarchy (Full, Main, Chap) and in both corpora (Osa-0r and Osa-1r). This fact means that we can improve the results for each granularity when no reduction or the 1% reduction is applied for every granularity, solely by transforming the single-task setting into a multi-task one leveraging the ICD hierarchically.

C. Towards Reproducible Research: Software Release

In an attempt to promote reproducible research, through this article we make available the software implemented, both single-task and multi-task models for multi-label classification developed for BERT.

The multi-label sequence classification head released with this paper is also compatible with Roberta [6], XLNet [43], XLM and DistilBERT [44]. The code is released and can be downloaded as follows:

- http://ixa2.si.ehu.eus/prosamed/MultitaskTransformers_soft
- with user: `MultitaskTransformers`
- and password: `IXAMultitaskTransformers`

Note that, whenever the the software is used in any way, this article should be cited in return.

VI. Conclusion

In this work, we have tackled the EHR multi-label classification problem according to the ICD in Spanish. Regarding the core methods, we applied state of the art Transformers (BERT Base Multilingual and BioBERT) to handle the high lexical variability, a distinctive characteristic of the biomedical field [45]. It is important to note that Big Data are not available in Spanish, hence, it is important to make the most of the training stage of the BERT base approaches both in the LM pre-training and in the fine-tuning of the multi-label classification module.

The aim of this work is twofold: on the one hand, to investigate if limited though in-house and in-domain data could help to further enrich LMs in such a way that the improvement in the underlying representation is noticed in our downstream task, i.e. the extreme multi-label classification. On the other hand, to find a method to exploit the hierarchical characteristics of the ICD that could help modelling the relationships between labels in this context with few annotated data. To this end, we conceived a strategy that, using the different granularities that can be extracted from an ICD label thanks to its hierarchy, enabled our multi-label models to become also multi-task learners i.e., the multi-label multi-task Transformer.

To test our hypotheses we designed two sets of experiments. From the first set, we assessed the impact of enhancing the LM. We learned that the performance of the BERT base models can be improved by further pre-training them with in-house unannotated data. This fact means that having access to limited, though, similar corpora can aid with the multi-label classification downstream task. The second set of experiments, based on multi-tasking, revealed leveraging the hierarchical fashion of the ICD approaching the ICD coding as a multi-task classification problem. We managed to boost ICD classification without the need of additional annotated data, benefiting from the created synergies between the different sets of labels. This finding is

especially relevant for downstream tasks with few resources, as it is the case of Spanish.

We would like to highlight that our multi-label multi-task Transformers, were built on top of the Huggingface Transformers library. To that end, we have developed a general multi-label multi-task head, ready to work with an arbitrary number of tasks. Moreover, there are no special requirements regarding the input (EHRs) to be suitable to run with this software. Any plain-text EHR written in any language can be used with no further modifications. Regarding the label-set, any EHR with the same version of the ICD is suitable. We are glad to announce that, with this article, we made available the proposed implementation of the hierarchical approach (details are given in Section V-C).

In this work we applied both BERT and BioBERT settings as they are well-known, well-maintained and widespread and, furthermore, extended general and biomedical model. Besides, BERT-based models count on small versions that are computationally affordable. In any case, our contribution, given that our software is built on top of the Huggingface Transformers library, could be adapted to any other transformation method (with Transformer-based architecture) seamlessly. That said, we do not rule out achieving better results with other transformation methods. Moreover, there is ample room for further progress in the extreme multi-label classification of EHRs according to the ICD. In one hand, the way towards incorporating more and more ICD labels could be further increased by incorporating zero-shot strategies [46] opening a way to the prediction of ICD codes only present on the test set. On the other hand, we found that getting unannotated data to enrich the language model is especially important for minority languages, nevertheless, finding further clinical EHRs or related corpora in languages other than English is the bottleneck [23], [24]. To bridge this gap, a promising strategy could be the unsupervised automatic translation [47] of majority into minority languages that could provide hundreds of thousands of EHRs, but at the expense of lower text quality. Along the same lines, abstractive summarizers [48] can be applied to generate alternative versions of the existing EHRs, doubling the number of available texts.

Acknowledgment

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] A. Névéol *et al.*, "CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French," in *Proc. CLEF (Work. Notes)*, 2017, pp. 1–17.

[2] A. Névéol *et al.*, "CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian." in *Proc. CLEF (Work. Notes)*, 2018, pp. 1–18.

[3] L. Goeuriot *et al.*, "Overview of the CLEF eHealth evaluation lab 2020," in *Proc. Int. Conf. Cross- Lang. Eval. Forum Eur. Lang*, Cham, Switzerland: Springer, 2020, pp. 255–271.

[4] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger, "Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020," in *Proc. Work. Notes Conf. Labs Eval. Forum. CEUR Workshop Proc.*, 2020, pp. 1–29.

[5] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[6] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[7] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[8] M. Peters *et al.*, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Computat. Linguistics: Hum. Lang. Technol.*, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Computat. Linguistics: Hum. Lang. Technol.*, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[10] H. Zhang and L. Xiao and Y. Wang, and Y. Jin, "A generalized recurrent neural architecture for text classification with multi-task learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3385–3391. [Online]. Available: https://doi.org/10.24963/ijcai.2017/473

[11] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, Doha, Qatar, Oct. 2014, pp. 1746–1751. [Online]. Available: https://www.aclweb.org/anthology/D14-1181

[12] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[13] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[14] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3615–3620.

[15] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proc. 18th BioNLP Workshop Shared Task*, Florence, Italy, D. Demner-Fushman, K. B. S. Cohen Ananiadou, and J. Tsujii, Eds., Association for Computational Linguistics, 2019, pp. 58–65.

[16] S. Amin, G. Neumann, K. Dunfield, A. Vechkaeva, K. A. Chapman, and M. K. Wixted, "MLT-DFKI at CLEF eHealth 2019: Multi-label classification of ICD-10 codes with BERT," in *Proc. CLEF (Work. Notes)*, 2019, pp. 1–15.

[17] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, and A. Bahamonde, "Binary relevance efficacy for multi-label classification," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 303–313, 2012.

[18] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[19] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[20] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[21] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2873–2879.

[22] X. Chen and C. Cardie, "Multinomial adversarial networks for multi-domain text classification," in *Proc. Conf. North Amer. Chapter Assoc. Computat. Linguistics: Hum. Lang. Technol.*, 2018, pp. 1226–1240.

[23] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than English: Opportunities and challenges," *J. Biomed. Semantics*, vol. 9, no. 1, pp. 1–13, 2018.

[24] H. Dalianis, *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Basingstoke, U.K.: Springer, 2018.

[25] M. Almagro, R. Martinez, V. Fresno, and S. Montalvo, "ICD-10 coding of spanish electronic discharge summaries: An extreme classification problem," *IEEE Access*, vol. 8, pp. 100073–100083, 2020, doi: 10.1109/ACCESS.2020.2997241.

[26] C. Mou and J. Ren, "Automated ICD-10 code assignment of nonstandard diagnoses via a two-stage framework," *Artif. Intell. Med.*, vol. 108, 2020, Art no. 101939.

[27] Z. Zhang, J. Liu, and N. Razavian, " BERT-XML: Large scale automated ICD coding using BERT pretraining," in *Proc. 3rd Clin. Natural Lang. Process. Workshop*, 2020, pp. 24–34.

[28] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, "Diagnosis code assignment: Models and evaluation metrics," *J. Amer. Med. Informat. Assoc.*, vol. 21, no. 2, pp. 231–237, 2014.

[29] A. Rios and R. Kavuluru, "Few-shot and zero-shot multi-label learning for structured label spaces," in *Proc. Conf. Empirical Methods Natural Lang. Process. Conf.*, ACM, 2018, vol. 2018, pp. pp. 3132–3142.

[30] W. Saib, T. Chiwewe, and E. Singh, "Hierarchical deep learning classification of unstructured pathology reports to automate ICD-O morphology grading," 2020, *arXiv:2009.00542*.

[31] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2020, *arXiv:2009.06732*.

[32] H. Su*et al.*, "Multilabel classification through structured output learning-methods and applications," Ph.D dissertation, Aalto Univ., 2015.

[33] J. B. Chrystal and S. Joseph, "Text mining and classification of product reviews using structured support vector machine," *Int. J. Multimedia Appl.*, vol. 5, no. 4, pp. 21–31, 2015.

[34] Y. Gu *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," 2020, *arXiv:2007.15779*.

[35] S. Gururangan *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8342–8360.

[36] W. H. Organization *et al.*, "International classification of diseases (ICD-10) world health organization," *International Classfication of Disease and Causes of Death. 9th Revision*. Geneva, Switzerland: WHO, 1975.

[37] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, and N. Taright, "Supervised topic models for diagnosis code assignment to discharge summaries," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2016, pp. 485–497.

[38] A. Blanco, A. Pérez, and A. Casillas, "Extreme multi-label ICD classification: Sensitivity to hospital service and time," *IEEE Access*, vol. 8, pp. 183 534–183545, 2020, doi: 10.1109/ACCESS.2020.3029429.

[39] Q. Yang and L. Shang, "Multi-task learning with bidirectional language models for text classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[40] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics - 18th China National Conference (Lecture Notes in Computer Science Series)*, Kunming, China, vol. 11856, M. Sun, X.Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham, Switzerland: Springer, 2019, pp. 194–206.

[41] Y. Huang, W. Wang, L. Wang, and T. Tan, "Multi-task deep neural network for multi-label learning," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 2897–2900.

[42] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[43] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, vol. 32, pp. 7059–7069. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf

[44] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, A distilled version of BERT: Smaller, faster, cheaper and lighter," *CoRR*, 2019, *arXiv:1910.01108*.

[45] K. B. Cohen and D. Demner-Fushman, *Biomedical Natural Language Processing*, vol. 11. Amsterdam, Netherlands: John Benjamins Publishing Company, 2014.

[46] R. Puri and B. Catanzaro, "Zero-shot text classification with generative language models," *CoRR*, 2019, *arXiv:1912.10165*.

[47] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *Proc. Int. Conf. Learn. Representations*, 2017, *arXiv:1710.11041*.

[48] A. Savelieva, B. Au-Yeung, and V. Ramani, "Abstractive summarization of spoken and written instructions with BERT," *CoRR*, 2020, *arXiv:2008.09676*.

On the Contribution of Per-ICD Attention Mechanisms to Classify Health Records in Languages With Fewer Resources than English

Alberto Blanco[1], Sonja Remmer[2,3], Alicia Pérez[1], Hercules Dalianis[2,3], and Arantza Casillas[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia, Spain
[2]Department of Computer and Systems Sciences, Stockholm University, Sweden
[3]Norwegian Centre for E-health Research, Tromsø, Norway
{*alberto.blanco,alicia.perez,arantza.casillas*}*@ehu.eus*
{*remmer,hercules*}*@dsv.su.se*

Abstract

We introduce a multi-label text classifier with per-label attention for the classification of Electronic Health Records according to the International Classification of Diseases. We apply the model on two Electronic Health Records datasets with Discharge Summaries in two languages with fewer resources than English, Spanish and Swedish. Our model leverages the BERT Multilingual model (specifically the Wikipedia, as the model have been trained with 104 languages, including Spanish and Swedish, with the largest Wikipedia dumps[1]) to share the language modelling capabilities across the languages. With the per-label attention, the model can compute the relevance of each word from the EHR towards the prediction of each label. For the experimental framework, we apply 157 labels from Chapter XI – Diseases of the Digestive System of the ICD, which makes the attention especially important as the model has to discriminate between similar diseases.

1 Introduction

Electronic Health Records (EHRs) are classified by clinical experts for documentation, reporting global health vital statistics, insurance billing, etc. International Classification of Diseases (ICD) is used world-wide to define diagnostic terms and procedures and serves to encode EHRs. There are thousands of terms encoded within the ICD WHO (2016). For medical experts, reading EHRs, lengthy and technical documents, finding explicit and implicit mentions of diagnoses and procedures for then assigning standard ICD codes is cumbersome and requires specific training. In fact, it is well-known that manual encoding is not error-free, as an example, Jacobsson and Serdén (2013), estimated that 20% of them were either incorrect or were missing. In this context, natural language understanding brings opportunities to bridge the needs of the society in terms of computer aided coding approaches.

In 2006, it was argued that Natural Language Processing (NLP) tools could quickly help identify codes in discharge summaries Kukafka et al. (2006). Today NLP tools for classifying clinical documents written in English are widespread. Even more, languages with scarce resources for biomedical NLP like Spanish, Italian, Swedish, etc., are in the limelight in the last years to develop codification systems as has been done for English. In the context of working towards the codification of documents, in languages with a small number of resources for NLP, different tasks have been addressed. In 2018 CLEF Névéol et al. (2018b) worked with Italian, French and Hungarian for the automatic codification of death certificates. Each death certificate consisted of a few words (on average 20 words) with at least one main diagnosis. In 2020 the CodiEsp task at CLEF Miranda-Escalada et al. (2020) consisted on the automatic assignment of ICD-10 codes to Spanish Clinical Records with 350 tokens on average. For Swedish Henriksson et al. (2011) the authors mentioned that the corpus was compiled with documents that on average had a length of 96 words.

Admittedly, multi-label classification is **challenging**, particularly with extensive label-sets (as it is the case of the ICD) and domain-specific corpora, and even more when it comes to dealing with clinical information extraction on languages other than English Névéol et al. (2018a). Spanish and Swedish researchers are striving to bridge this gap, indeed, as the first and relevant step, they gathered corpora conveying patient records Oronoz et al. (2015); Dalianis (2018). Previous works showed that the multi-label classification problem of EHRs coded with ICD-10 can be tackled with an adapted

[1]https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages

Proceedings of Recent Advances in Natural Language Processing, pages 165–172
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_020

BERT architecture Amin et al. (2019); Zhang et al. (2020).

Moreover, we focused just on a sub-set of the ICD, i.e. the Diseases of the Digestive System (the ICD codes starting with the letter K). Focusing on semantically related diseases poses an added challenge, since the Natural Language Understanding (NLU) in charge of encoding the input EHR must be able to cope with the nuances inherent to the distinction of similar diseases. Unarguably, it is easier to distinguish two diseases each belonging to a different body-part than two diseases within the same body-part (as it is this case distinguishing diseases all within the digestive system). In summary, distinguishing semantically different diseases (e.g. gastrointestinal vs cardio-pulmonary) would be easier than distinguishing two diseases within the same speciality. To that end, the LM and the attention mechanisms play the most critical role, so we opted for the transformers models. BERT-based approaches have been tested in this context, with attention mechanisms as a strength towards finding relationships between input text with output ICD codes. The attention is a mechanism whose effectiveness has also been shown with other architectures such as RNNs with LSTM units Hochreiter and Schmidhuber (1997) or Convolutional Neural Networks Du et al. (2017).

Nevertheless, in this context we are dealing with scarce resources and relatively similar codes. In this line, the main scientific **contribution** of this paper rests on the implementation of a head adapted for BERT with multiple label attention mechanisms (instead of a generic one) in order to delve deeper into the nuances of the understanding module. In this work, we have implemented a per-label attention mechanism, and given that regular BERT models also have the self-attention mechanism, it allowed us to compare the effect of different attention mechanisms. The per-label attention mechanism allows the model to give a different relevance to each word and ICD code pair, contrary to the regular attention mechanism. The experimental results support the approach's acceptable performance, so we decided to release the head for the scientific community.

2 Corpora

We have applied two datasets of languages with scarce resources for this work, i.e., languages with fewer resources than English, specifically, Spanish and Swedish. Both datasets are Electronic Health Records containing Discharge Summaries from patients. The Spanish EHRs are from the Emergency Services of the Basque Health Public System, conveying records, and therefore labels, from all the medical specialities Oronoz et al. (2015). However, the Swedish EHRs are only from the gastrosurgery medical specialisation and comes from the research infrastructure Health Bank - Swedish Health Record Research Bank[2], at Stockholm University. Therefore, to have equal label sets, we have selected the ICD codes shared between both datasets to carry out the experiments, obtaining 157 codes, all from the Chapter XI of the ICD-10, i.e., Diseases of the Digestive System. By selecting the codes of some specialities the number of available EHRs is reduced but the label sets are easier to handle. Training specific models on EHRs of specialities improves the performance against training general models Blanco et al. (2020). For the Swedish ICD-10 corpus data set the Swedish KB-BERT model Malmsten et al. (2020) has been applied with good results, see (Remmer et al., 2021).

Here we present a quantitative description and comparison between both datasets. Regarding the input, the Swedish dataset is more than twice larger in number of EHRs, with 8,909 records in contrast to the 3,891 available Spanish EHRs. Nevertheless, the vocabulary (i.e., number of unique words) is around three times bigger for the Spanish dataset. One explanation is that the Spanish EHRs come from several specialities, and therefore there is a higher lexical variability due to the specific terms of each medical specialisation. Also, the Spanish EHR contains lab tests, which could increase the number of unique words significantly.

Regarding the output, both datasets are equivalent, with the same set of 157 gastrointestinal ICD-10 codes. Although this is just a subset of the labels, there are still infrequent codes. For example, only 45 codes from the 157 appear in at least 1% of the EHRs. This fact makes the task even more challenging, as, for around 28% of the labels, there are only a few samples from where the model can learn. Even though the number of labels is the same, the distinct label sets (i.e., label combinations that are unique) are larger in the Swedish dataset than in the Spanish (1,288 and 558, respectively) due to the higher number of records. The ratio between the distinct label sets and the number of records

[2] http://dsv.su.se/healthbank

is similar, 6.97 for Spanish and 6.91 for Swedish, meaning that about the same number of EHRs lead to the same number of unique label sets.

The most significant differences come when evaluating the length of the EHRs, as the Spanish EHRs are significantly longer. While the Spanish records convey 984 words on average, the Swedish only have 74 words. The standard deviation is also more prominent in proportion, with 491 for the Spanish and 77 for the Swedish (note that the standard deviation is higher than the mean). Although the records from both datasets are Discharge Summaries, it seems that not all the Swedish records are complete summaries, but instead a summary or even one-sentence synopsis of the patient's outcome.

3 Methodological Approach

Focusing the attention on the methodology, in Amin et al. (2019) the authors demonstrate the effectiveness of transfer learning with pre-trained language representation model BERT without attention for the multi-label classification of German non-technical summaries (NTSs) of animal experiments. In e-Health 2020 the authors of López-Garcıa et al. (2020) tackled the task as a multi-label classification problem using BERT model Devlin et al. (2019) for the automatic clinical coding of medical cases in Spanish. NLU results crucial to this task and Transformers-based Language Models (LM) are, doubtlessly, the key strength of most recent approaches such as multi-label biomedical text classification Gu et al. (2020). All this and the inherent challenges related to our work (e.g. the ability to distinguish concepts leveraging semantically related diseases) **motivated** us towards BERT-based approaches. Another fact in favour to this choice rests on the ability to the transfer learning between the two languages and, if possible, get benefits from one Language Model to the other. That is, the resources from one language can boost the LM of the other one, while the system remains decoupled from the data.

In order to tackle the multi-label text classification task, we applied a model with a Transformer-based architecture. The problem to solve is the mapping between the input of the EHRs (the raw text, X) and a subset of ICDs from the entire label set, $\mathcal{C}$, where $|\mathcal{C}|$ is the number of codes. The Deep Learning model is trained for the downstream task with pairs of input and output (i.e., EHR texts and ICD codes). The Transformer-based neural network model is trained with instances comprising pairs of input (EHR text) and output (ICD codes). The j-th instance is described formally as $(X_j, \mathbf{c}^j) \subseteq \Sigma^* \times \{0,1\}^{|\mathcal{C}|}$. The input-output pair is as follows: X_j is the string of any length (comprised by tokens from the vocabulary Σ), i.e. the EHR. $\mathbf{c}^j$ is a presence-bits array. $\mathbf{c}_i^j$ encodes the absence or presence of each code $C_i \in \mathcal{C}$ linked to the instance X_j, i.e. the ICDs assigned to the EHR.

From the input text, X_j, fed to the Language Model part of the model, a hidden document representation is obtained. The importance of this rests in that our multi-label classifier is built on top of a BERT model (see Section 3.1). The LM is the core of the Transformer-based NLP models. The principal contribution of this work is the use of the hidden representation to compute attention weights that are label-specific for each input token. After computing the attention, the final output (label predictions) is computed with a fully connected layer that is fed with another document representation got from the label-specific attention layers. To support the reproducible research, we release the code of the per-label attention mechanism with this article.

3.1 Baseline: BERT to Boost LM

The Language Models based on Transformers, specifically BERT models Devlin et al. (2019), have been acknowledged due to their ability to generate contextual representations. In this work, we have to differentiate between very similar diagnoses (all from the gastrointestinal service), which motivated the chosen BERT model as the LM part of our multi-label text classification system to generate the representation of the EHRs. A BERT model is also suitable because of its built-in self-attention function, which can connect different locations of a single input sequence to one another. We also turned to BERT because it has been shown to expand Recurrent Neural Networks' ability to model dependencies to long-distance patterns Hochreiter and Schmidhuber (1997).

In an attempt to encompass Spanish and Swedish, EHRs were represented with shared LMs. The transfer learning approach of sharing the LM poses two advantages. On the one hand, it alleviates the training process for each language since just the task-dependent module (i.e. ICD multi-

label classification) has to be trained. On the other hand, this bypasses the lack of in-domain data for languages other than English. Indeed, the multi-lingual LM, with English, leverages other languages such as Spanish and Swedish in a synergistic effect since cross-language regularities are captured Pires et al. (2019).

The LM part is the core of the BERT models, but coupling different heads on top of the LM is what concedes the ability to tackle numerous downstream tasks, as multi-label classification. Since there are many parameters to describe both the LM and the head for the downstream task, training a BERT model is challenging. The LM module contains the broad majority of the parameters that must be inferred during the training stage. The ICD multi-label classification head built for this study, for example, accounts for less than 1% of the total model parameters (even though using the smallest variant of BERT, which has $110M$ of parameters). With this in mind, we opted to train the multi-label heads from scratch while fine-tuning the LMs instead of training the LMs from scratch.

Because of memory and computational limitations, we used the $\text{BERT}_{\textbf{BASE}}$ as the baseline BERT model (our GPUs are limited to 8GB of DRAM memory). The $\text{BERT}_{\textbf{BASE}}$ model comprises 12 Transformers blocks, 12 self-attention heads, and an internal embedding layer size (d) of 768, totaling $110M$ parameters. The pre-trained $\text{BERT}_{\textbf{BASE}}$ Multilingual model was used. The downstream tasks' attention and output layers are connected to the output of LM, the hidden document representation, ($\mathbf{H}$), of the EHR.

3.2 Contribution: Per-ICD Attention Head

Having opted for the multilingual BERT to cope with the LM, next we proposed to improve the task-dependant head. The aim was to leverage ICD-dependant attention mechanisms in an attempt to enhance the model with added NLU capability when it comes to distinguishing ICDs within the same hospital-service (Digestive in our case).

Our multi-label classification head incorporates a per-label (per-ICD) attention mechanism. The model can classify the EHRs with respect to the ICD labels that are present through the text while also calculating the importance that each input token (word) has in relation to each of the ICDs.

Here, N is the number of tokens of the EHR (length) and d is the BERT hidden layer dimension (i.e., the representation of documents, being $d = 768$ for $\text{BERT}_{\textbf{BASE}}$ models). Then, rather than perform the pool operation (across the document length, N), as in the original BERT Devlin et al. (2019) for classification, our head uses a per-ICD attention mechanism. The per-ICD attention mechanism allows the classifier to discover the correct relationships between the input tokens and each label.

For each ICD label, C_i, the attention vector $\boldsymbol{\alpha}_{C_i} \in \mathbb{R}^{|C|\times N}$ is computed from the learnable vector parameter $\mathbf{u}_{C_i} \in \mathbb{R}^d$, following (1), where C is the full set of ICD labels.

$$\boldsymbol{\alpha}_{C_i} = Softmax(\mathbf{H}^T \mathbf{u}_{C_i}) \tag{1}$$

The attention scores must be computed as a probability distribution, representing the importance between each token and ICD label pair, and to that end, the model leverages the Softmax function. The matrix multiplication between $\boldsymbol{\alpha}$ and $\mathbf{H}$ is calculated to get an ICD representation for each class from the attention weights. In the end, the maximum through the labels' dimension is taken, obtaining the document representation on the final layer ($\mathbf{v} \in \mathbb{R}^d$), which combines the per-ICD attention representation.

The final layer of the head for multi-label classification is a regular one that allows getting the probabilities for each ICD label. It is a linear layer that takes the document representation ($\mathbf{v}$) as input, which takes into account the attention weights for each input token and label pair. After that, a Sigmoid function is applied to get the actual probabilities of each ICD, as in (2).

$$\hat{\mathbf{y}}_\mathbf{i} = \sigma(\mathbf{W}_\mathbf{i} v_i + b_i) \tag{2}$$

The probability of each ICD class ($C_i \in C$) being on the given input text is $\hat{y}_i$. The parameters of the final layer are the weights matrix ($\mathbf{W}$) and bias ($\mathbf{b}$). Regarding the training of the model, it is carried out by minimising the loss function, precisely, the Binary Cross-Entropy (BCE) loss, as in (3). On this equation, the $\hat{\mathbf{y}}$ is the output of the previous final layer, and $\mathbf{y}$ is the vector that encloses the ICD codes present on the EHR (i.e., the appearance or lack of ICD codes). Figure 1 shows an architectural outline of the system.

$$BCE(\hat{\mathbf{y}}, \mathbf{y}) = -W[\mathbf{y}\log(\hat{\mathbf{y}}) + (1-\mathbf{y})\log(1-\hat{\mathbf{y}})] \tag{3}$$

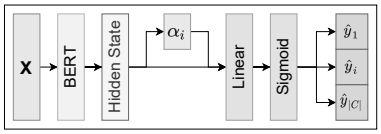


Figure 1: Architectural outline of the developed per-ICD BERT model

4 Experimental Framework

We propose the following experimental setup to evaluate our BERT model's performance with per-ICD attention compared to the benchmark (standard BERT model) on the multi-label ICD classification downstream task. The experimental setup comprises the two minority languages (in terms of in-domain clinical data available), Spanish and Swedish, and a gastrointestinal label set of 157 labels. Each experiment is carried out twice, with the same experimental and training parameters, one with the regular multi-label classification head (as the baseline) and the other with our head with per-ICD attention. We show the results from the experimental results in Table 1 and Figures 2 and 3, for Spanish and Swedish, respectively.

The model with our per-ICD attention head obtains better results in both languages. It is important to note that the results improve considerably even in this context with a considerably large label set (157 labels). This finding is consistent with the following hypothesis: many terms can be important when dealing with a wide number of ICD codes at once and long EHRs, but probably only a few of them are relevant for each ICD code individually.

Multi-label ICD classification is often assessed by means of the Area Under the ROC Curve (AUC) micro averaging the metric for all the ICDs involved (denoted as AUCm in Table 1). For Spanish, the per-ICD model surpasses the base BERT model by 9.16 points, also improves slightly for the Swedish, with an improvement of around 1 point. In Figures 2 and 3 we show the confusion matrices for each experiment. Each confusion matrix is the average of the matrices of each ICD class, and we have computed two versions, i.e. one with arithmetic averaging (aka samples average) and the other with weighted averaging. In both, the darker the colour, the higher the metric, always in the range $[0 - 100]$. The weighted averaged matrices are computed considering the support (relative frequency) of each ICD class. Note that the TPR (True Positive Rate) and FNR (False Negative Rate) shown in Table 1 are also the arithmetic average of each corresponding model, but the CM show also the FPR (False Positive Rate) and TNR (True Negative Rate), while the weighted average of each metric. Regarding the per-class performance, there is a positive association with the support; the more frequent the label, the better are the results.

If we analyse the matrices, it can be observed that the source of improvement of the per-ICD model can be broken down; while the True Negatives stay close (as with a large label set, the majority of classes are negative), the True Positives improves considerably, with an increment of almost 100%. In the same way, the False Negatives decrease by around 20%. Although the Swedish results are in general weaker, this behaviour is appreciated similarly for both languages. Therefore, given the results, it seems that our per-ICD attention head is able to improve the Precision of the regular BERT models for ICD multi-label classification with large label sets. Nevertheless, the per-ICD model outperforms regular BERT in terms of performance, but also in interpretability capabilities, as it has the ability to export the attention weights, allowing its visualisation.

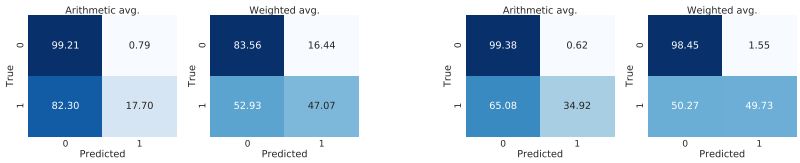
L	Model	AUCm	TPR	FNR
SP	baseline	58.16	17.70	99.21
	per-ICD	**67.32**	**34.92**	**99.38**
SW	baseline	54.92	15.49	92.24
	per-ICD	**55.96**	**27.91**	**82.45**

Table 1: Comparison of results on the Spanish (SP) and Swedish (SW) datasets ("L" stands for "Language") obtained with the baseline BERT and BERT enhanced with per-ICD attention head. TPR is the True Positive Rate and FNR the False Negative Rate.

5 Discussion

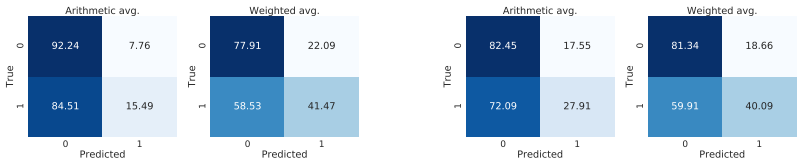
Within the clinical text mining field, the main weakness tends to be the availability of corpora due to the natural patient's confidentiality policy Cohen and Demner-Fushman (2014). As a result, for the research to make progress, the so important comparability might get compromised. By contrast, through this work the authors are glad to make available their own implementation of the per-ICD attention approach[3] as a secondary contribution of

[3]To get the source code of the implementation, simply e-mail the first author.



(a) Baseline BERT (b) Per-ICD BERT

Figure 2: Average heatmaps of Spanish models



(a) Baseline BERT (b) Per-ICD BERT

Figure 3: Average heatmaps of Swedish models

this paper.

Another aspect related with the corpus is the complexity and length of the input EHR. The average length of the input of the works mentioned Névéol et al. (2018b); Cappellato et al. (2019) are variable from a few words in the case of Italian, Hungarian and French to 350 words for the documents written in Spanish Miranda-Escalada et al. (2020). By contrast, in our paper we deal with documents in Spanish and Swedish with an average length of 800 (exceeding the aforementioned ones) and 70 respectively.

According to these results, the per-label attention mechanism improves Precision. While more performance is still necessary for a fully automated system, the results suggest that it is suitable for multi-label classification of EHRs according to the ICD standard, specifically applying it as a clinical DSS, as the per-ICD attention can aid the expert in the EHR codification process.

6 Conclusions

We have dealt with the codification of EHRs of the gastrointestinal service for Swedish and Spanish hospitals. We have developed a BERT model for multi-label classification incorporating a per-label attention mechanism.

The results obtained have revealed that the proposed model outperforms the regular BERT. We have proved this fact for two languages with minority resources in clinical NLP, showing that solutions of language independent nature work. Moreover our proposal generates an interpretable output that helps to know the relevance of the tokens with respect to each ICD assigned to the EHR. To sum up, the per-label attention mechanism differentiates semantically ICDs that are related and aids to explain the core of each label. Future work may include testing BERT models trained for the specific languages, as the BETO model Cañete et al. (2020) for Spanish.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), European Commission (FEDER) and the Basque Government (IXA IT-1343-19, Predoctoral Grant PRE-2019-1-0158) and by the ClinCode project, project number 318098, from the Norwegian Research Council. This research has been approved by the Regional Ethical Review Board in Stockholm under permission no. 2007/1625-31/5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. 2019. MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In *CLEF (Working Notes)*, pages 1–15.

Alberto Blanco, Alicia Pérez, and Arantza Casillas. 2020. Extreme multi-label icd classification: Sensitivity to hospital service and time. *IEEE Access*, 8:183534–183545.

Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors. 2019. *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Practical ML for Developing Countries at ICLR 2020*.

Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, pages 4171–4186.

Jiachen Du, Lin Gui, Ruifeng Xu, and Yulan He. 2017. A convolutional attention model for text classification. In *National CCF conference on natural language processing and Chinese computing*, pages 183–195. Springer.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Aron Henriksson, Martin Hassel, and Maria Kvist. 2011. Diagnosis code assignment support using random indexing of patient records – a qualitative feasibility study. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 348–352. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Anders Jacobsson and Lisbeth Serdén. 2013. Kodningskvalitet i patientregistret, (In Swedish).

Rita Kukafka, Michael E Bales, Ann Burkhardt, and Carol Friedman. 2006. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association*, 13(5):508–515.

Guillermo López-Garcıa, José M Jerez, and Francisco J Veredas. 2020. ICB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020.*

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018a. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13.

Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. 2018b. CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In *CLEF (Working Notes)*, pages 1–18.

Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza, and Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sonja Remmer, Anastastios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. *In the Proceedings of Recent Advances in Natural Language Processing, RANLP 2021, Varna, Bulgaria*.

WHO. 2016. International Classification of Diseases (ICD). Accessed 2021-04-14.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. *arXiv preprint arXiv:2006.03685*.