Expert Systems With Applications 204 (2022) 117303

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Preliminary exploration of topic modelling representations for Electronic Health Records coding according to the International Classification of Diseases in Spanish

Nuria Lebeña[1], Alberto Blanco [*,1], Alicia Pérez[1], Arantza Casillas[1]

HiTZ Center - Ixa, University of the Basque Country (UPV/EHU), Manuel Lardizabal 1, 20080 Donostia, Spain

ARTICLE INFO

Keywords:
Multi-label classification
Document classification
Electronic Health Records
ICD classification
Topic models
Partially labelled dirichlet allocation

ABSTRACT

In this work, we cope with the classification of Electronic Health Records (EHR) in Spanish according to the International Classification of Diseases (ICD). We employ Topic Models representing each document as a probabilistic distribution over topics, offering a low-dimensional representation of documents.

The trend is to turn to an embedding text representation, but these approaches require large amounts of textual data. We found Topic Models as a suitable alternative approach to deal with the few resources available for Spanish clinical text mining. Besides, they are interpretable and aid the explainability in artificial intelligence (XAI).

We explored two different methods, known as Latent Dirichlet Allocation (LDA) and Partially Labelled Latent Dirichlet Allocation (PLDA), the supervised approach of the former. We assessed the results attained in Spanish with an analogous task in English as a reference. Evaluation methods were applied directly to the representation, with metrics to determine topic coherence and the relationship between topics and ICD labels.

We learned that PLDA was able to discover topics associated with the ICD. This finding means that this representation itself can reveal ICD codes previous to classification. Also, this representation was used as predictive features to feed a conventional classifier to show their competence in a downstream task. We conclude that in a context with a lack of big data availability, PLDA emerges as a versatile candidate, able to offer a competitive representation of EHRs.

While other works are primarily concerned with supervised categorization and do not pay attention to the representation, LDA and PLDA offer an interpretable approach that can be associated with ICDs. Moreover, compared with those that employ LDA, we demonstrate how its' supervised version, PLDA, can be more intuitive as it shows a closer relation with the ICDs.

1. Introduction

Electronic Health Records (EHR) are documents written by doctors that describe patients' antecedents, chief complaint, hospital course, discharge medications etc. EHRs convey a vast and valuable amount of unstructured information, as an example, the EHRs in our datasets convey, on average, above 1000 tokens. Expert coders are in charge of reading and understanding the text, carefully seek procedures and diagnoses and thus, assign, the corresponding code following the International Classification of Diseases (ICD). Each EHR has multiple ICD codes assigned that help to summarize the content of the EHR as if they were keywords. The ICD comprises thousands of codes each of which is associated to a standard term that represents medical diagnoses or procedures. ICD codes serve for information sharing world-wide, enable the search of similar EHRs, help to extract statistics about diseases and mortality and are also used by insurance companies for billing.

Assigning codes to EHRs seems appropriate for Natural Language Understanding (NLU) since complex texts are summarized as a subset of specific terms. Indeed, EHR coding received the attention of international research challenges such as CLEF eHealth (Dörendahl, Leich, Hummel, Schönfelder, & Grune, 2019; Goeuriot, et al., 2020). ICD coding is not straightforward, since the EHRs rarely convey the terms expressed in the standard terminology (e.g. only a tiny fraction of the codes are found with an exact match Miranda-Escalada, Gonzalez-Agirre, Armengol-Estapé, & Krallinger, 2020). Some tasks, as CLEF

* Corresponding author.
E-mail addresses: nlebena001@ikasle.ehu.eus (N. Lebeña), alberto.blanco@ehu.es (A. Blanco), alicia.perez@ehu.eus (A. Pérez), arantza.casillas@ehu.eus (A. Casillas).

[1] All the authors have contributed equally.

https://doi.org/10.1016/j.eswa.2022.117303
Received 30 April 2021; Received in revised form 14 March 2022; Accepted 22 April 2022
Available online 6 May 2022

eHealth 2018 (Névéol, Robert, et al., 2018) deal with the classification of short fragments of EHRs with just diagnostic terms expressed in a non-standard manner (Névéol, Robert, et al., 2018). Instead, we cope with the entire document, since the diagnostic terms are not identified in unstructured EHRs.

According to the related work, the principal trend to assign one or more labels to long documents is to employ deep neural classifiers. The advantage of these approaches is that are able to handle the text encoding and the text classification itself (Devlin, Chang, Lee, & Toutanova, 2019; Vaswani, et al., 2017). However, while deep neural approaches can discover inherent features, these features are difficult to analyse by human experts and provide us with little feedback to keep making progress. Besides, it is well known that these approaches attain good performance given big data.

Nowadays it is easy to find free medical corpus written in English, for example, MIMIC (Johnson, et al., 2016), I2B2 (Uzuner, South, Shen, & DuVall, 2011) and the TREC (Voorhees & Hersh, 2012) corpus. Also, free available English medical embeddings such as Chen, Peng, and Lu (2019) and Kalyan and Sangeetha (2020) do exist. Generating embeddings is a task that demands large amounts of documents. Working with minority languages or tasks with few resources, as it is the case of Spanish clinical text mining, is not a favourable scenario to create good quality embeddings.

In this work we explored EHR encoding as a topic-modelling task (Dermouche, et al., 2016; Gangavarapu, Jayasimha, Krishnan, & S., 2020). This means that out of the main trend, that focuses on the use of classifiers to predict ICDs, we wondered if latent topics generated by the topic models representations would be representative of ICD codes without using a classifier. Topic models, as Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and forthcoming improved approaches (Blei, Ng, & Jordan, 2003; Hofmann, 2001; Wang, Bai, Stanton, Chen, & Chang, 2009), follow the intuition of summarizing a text with keywords. Indeed, topic models represent a document as a distribution of discovered topics: the presence of each topic in the document is accounted through a probability (as depicted in Fig. 1(a)). In their turn, each topic is represented as a distribution over the vocabulary: the words that best represent the topic get bigger weight in the distribution (as in Fig. 1(b)). While humans would directly summarize the EHR with a set of ICDs (nominal values), inferred topic models offer a distribution (quantitative values) over abstract topics described as a mixture over the vocabulary. Moreover, topics can be easily visualized in a human-friendly understandable manner (Chaney & Blei, 2012) enhancing natural language understanding with explainability. Indeed, the resulting vector representation can help find related documents with similarity metrics accounted for on the topic space.

We **hypothesize** that inferred topic models can serve, in absence of big data, to represent documents naturally following the ICD coding rationale. Our concern is to what extent are inferred topics related to the standard ICD labels. Provided that the representation resulted to be robust for EHR coding with respect to ICD, then it would be of much benefit due to the fact that topic-models provide a human-friendly representation that can boost NLU and the development of EHR classification.

In brief, this work deals with EHR encoding with the focus on explainable representations in the context of a task without big data. To this end we turned to two types of topic models: Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Partially Labelled Latent Dirichlet Allocation (PLDA) (Ramage, Manning, & Dumais, 2011). We assessed the models on related and comparable corpora, Osa in Spanish and MIMIC-III English (Johnson, et al., 2016). There is a difference, since Osa is encoded with ICD-10 and MIMIC-III with ICD-9. We assessed the topic models directly, by their coherence and by their correlation with respect to ICD, and also indirectly, for the information rendered to a simple classifier to predict the ICDs. We insist on the fact that our main interest rests on the ability of the representation to explain the content of the EHRs. This is why in this preliminary work we are not interested in complex classification approaches.

Table 1
An example of standard diagnostic terms (DT) in the corpus and their corresponding ICD codes. MIMIC-III corpus (in English) conveys ICD-9 and Osa (Spanish) ICD-10. L1, L2, L3 and L4 are representative of different specialties: cardiology, neurology and endocrinology.

	DT	ICD-9	ICD-10
L1	Diabetes mellitus W/o mention complication	250.0	E11.9
L2	Congestive heart failure	428.0	I150.9
L3	Essential hypertension	401.9	I10
L4	Acute kidney failure	584.9	N179.9

Table 2
Label-set frequency of the subset of labels in the example (Table 1) the frequency in English and Spanish differ notably. L1, L2, L3 and L4 are representative of different specialties: cardiology, neurology and endocrinology.

	L1	L2	L3	L4
English	1 416	2 114	3 231	1 447
Spanish	3 964	2 783	7 698	1 120

2. Materials

This work attempts at finding a representation of the biomedical text of the EHRs that could be applied within the documentation services of hospitals and health centres. To this end, we used an anonymized set of discharge records from Osakidetza, the public health system from the Basque Country. The EHRs are in Spanish and follow the ICD-10 classification. The methods applied to Spanish were compared with a similar task to English as a reference. That is, to encode EHRs according to ICD-9, and was developed employing the MIMIC-III corpus (Johnson, et al., 2016).

In an attempt to make both Spanish and English tasks comparable with previous work, in this preliminary work, we focused on labels appearing in at least the 1% of the samples following the trend (Blanco, Pérez, & Casillas, 2020; Mullenbach, Wiegreffe, Duke, Sun, & Eisenstein, 2018).

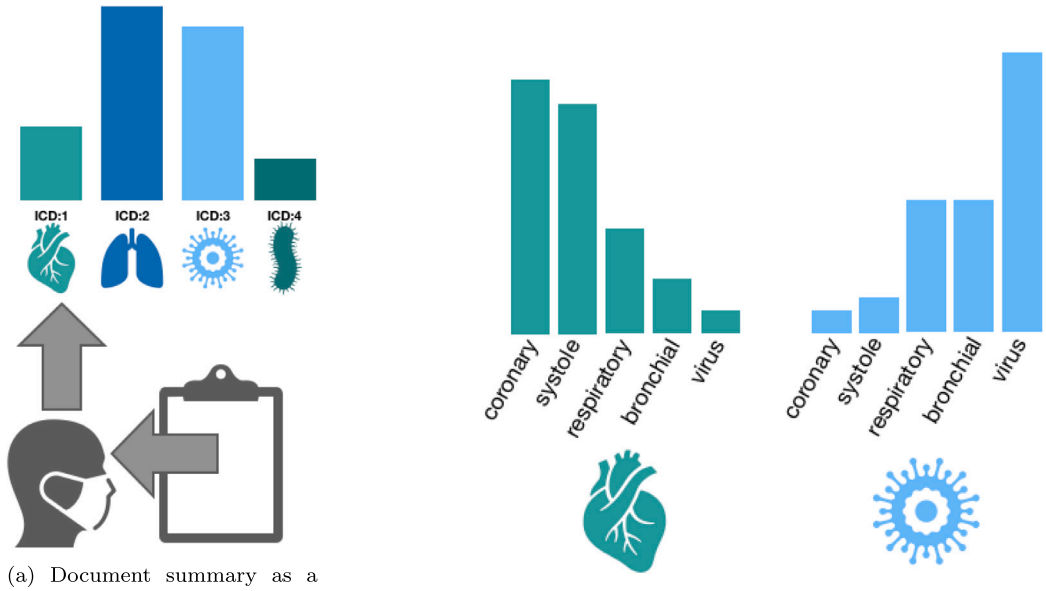
Fig. 2 shows the amount of labels per document in the 50 labels set in English and 104 in Spanish. Note that the ICDs differ, due to the fact that English and Spanish corpora are encoded, respectively with 9th and 10th revisions of the ICD. An example is provided in Table 1 to show explicitly the different versions of the ICD involved corresponding to a given Diagnostic Term. This example shows a small subset of diagnostic terms (DTs) that is frequently seen in elder patients. DT in both are equal, however the ICD version differs. The frequency of those labels is given in Table 2.

Note that the term "label", "class" or "ICD" is used interchangeably, as the label refers to the Diagnostic Term (DT), and each DT is encoded using an ICD, with the 9th version in English, or the 10th version in Spanish. To ease the reading and prevent redundancy, we renamed each pair of labels as L1–L4, as detailed in Table 2.

The corpora was randomly split following the iterative stratification approach (Sechidis, Tsoumakas, & Vlahavas, 2011). This approach produces splits in which the proportion of the labels are maintained. The resulting partition is shown in Table 3. Note that the size of the vocabulary is larger in the Spanish set. However, the tasks are comparable in volume of data in terms of both EHRs and ICDs. The mean of the quantity of labels per EHR is unbalanced, 5.7 in MIMIC-III and 3.86 in Osa. Nevertheless, the label-set distribution is similar as shown in Fig. 2, appearing the majority of the labels in almost 500 documents.

3. Methods

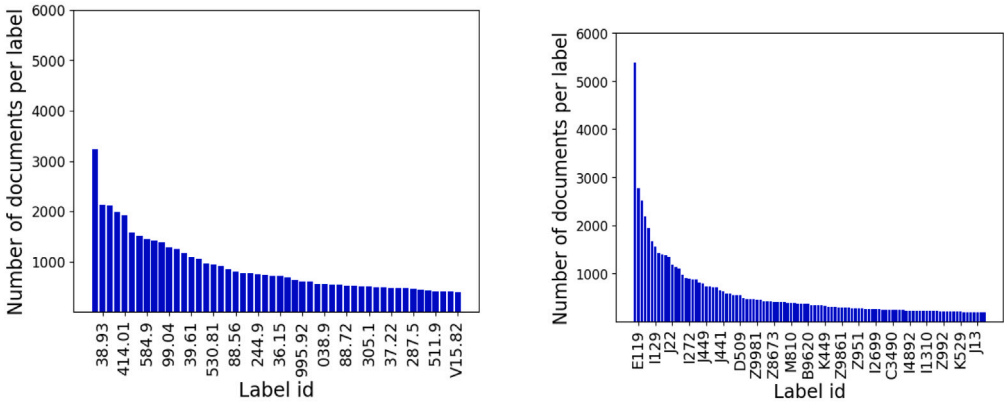
In this section, first, we introduce the topic models and compare with other document representations. Next, we describe different techniques used to evaluate the performance of these models.



(a) Document summary as a distribution over topics

(b) Topic definition as a distribution over the vocabulary

Fig. 1. Topic models defined as distributions.



(a) English

(b) Spanish

Fig. 2. Label-set distribution on Osa corpus in Spanish (encoded with ICD-10) and on MIMIC-III corpus in English (encoded with ICD-9).

Table 3
Corpus size for both Osa and MIMIC-III: the number of EHRs, the average number of tokens per EHR, the vocabulary found in the EHRs.

	Spanish		English	
	Train	Test	Train	Test
EHRs	15 791	6 820	8 059	3 460
Tokens	423	464	799	815
Vocab	56 277	38 022	41 496	31 981

3.1. Representation

Natural Language Understanding (NLU) relies on compact representations of text into numeric vectors (Liu, He, Chen, & Gao, 2019). It is crucial that vector representation enables a means of finding semantic relatedness between documents. Getting abstract representation of texts has evolved rapidly in last years.

Classical Bag of Words (BoW) representation merely indicated the presence of words from the vocabulary of the task within a document (Abdulaziz, M. Ameen, & Ahmed, 2019). This representation assumes that documents with similar patterns of presence/absence of words would convey similar information. The drawback of this representation is twofold: on the one hand, the dimension of the representation is equal to the size of the vocabulary, thus, tends to be of several thousands; on the other hand, BoW disregards word-ordering, thus, semantic patterns are out of the scope of BoW. Bag of n-grams arose as an alternative to capture low-range word-ordering, however, this happens at the expense of increasing the dimension of the space and, hence, is detrimental to the computation cost of inference algorithms.

Word and document embedding (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) emerged as the alternative to BoW offering a representation in a low-dimensional space in which semantically similar tokens are closely positioned as they have similar representations. In contrast to BoW, word embeddings are represented by a real-valued vector of tens or hundreds of dimensions, having a great performance when representing documents with large vocabulary. The drawback of word embedding rests on the fact that the representation is not easily interpretable by humans. That is, each coordinate and also the value of each coordinate is far from trivial. A recent study (Dieng, Ruiz, & Blei, 2020) has developed *Embedded Topic Models* (ETM) a generative model of documents that blends *topic models* with word embeddings. To acquire embeddings, massive volumes of domain-specific corpus are

required, as explained in Dieng et al. (2020); nevertheless, this is a significant difficulty for specific domains and languages (AlShuweihi, Salloum, & Shaalan, 2021; Névéol, Dalianis, et al., 2018). As it is in our case with unrealistic Spanish EHR corpora.

The motivation of this work is to get a well-performing representation of text for low-resourced languages or tasks. Word embeddings work well with languages with lots of corpora available. Without the need of extensive resources, topic models (Blei, 2012) offer a compact low-dimensional though interpretable representation of documents. The representation discovers latent semantic structure, indeed, the topics inherent to the document. Topic models represent each document as a mixture of topics, being each topic a mixture of terms (somehow depicted in Fig. 1). The dimension remains comparable to the word embeddings with the interpretability as an added value. Each coordinate refers to a topic and the higher the value the more intense the presence of the topic in the document. This reminds of document to keyword summarization rationale, and conveys a complex natural language understanding ability.

Topic models embrace the idea of class-based stochastic language models (LM) (Brown, Della Pietra, Desouza, Lai, & Mercer, 1992), with the topics involved seen as coarse-grained classes. Indeed, the joint probability of a word w in a document d, $p(d,w)$ is estimated involving the topics, t, as a hidden variable (the same role as the classes in LMs) and also assuming that words are drawn from topics with independence of documents as in (1).

$$\begin{aligned} p(w,d) &= \sum_t p(w,t,d) \\ &= \sum_t p(t)p(d|t)p(w|t,d) \\ &= \sum_t p(t)p(d|t)p(w|t) \end{aligned} \tag{1}$$

All the probability distributions are inferred from the corpus. The topics in the document, $p(t|d)$, and the words in the topic, $p(w|t)$, are modelled as Dirichlet distributions. These parameters are inferred through the Expectation Maximization algorithm. A Dirichlet distribution is used to randomly draw topics from documents and, in the inference process, model the topic mixture. Another Dirichlet distribution is involved to model the words in the topics. This idea led to Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is an unsupervised approach, latent topic-structure is inferred directly from a set of documents without making use of document labels.

Partially Labelled Latent Dirichlet Allocation (PLDA) (Ramage et al., 2011) is a supervised extension of LDA that builds topic statistical structure taking document labels into account. The topics in PLDA are bound to a given label. While LDA is able to reduce the dimensions of a set of document-to-term representation, PLDA is built focusing on subsets of documents with a given label, as a result a mixture of topics per label is obtained. Indeed, the goal is to get a mixture for each label, this time, a Gaussian model drives the approach as a continuous class distribution.

3.2. Normalized pointwise mutual information

Normalized pointwise mutual information is a metric aimed at assessing the interpretability of a topic (Rubin, Chambers, Smyth, & Steyvers, 2012) and it has been used to compare the quality of the model itself (Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012). The rationale is that interpretable topics tend to convey words with close meaning.

To begin with the relatedness of each pair of words (w_i, w_j) is estimated using the normalized point-wise mutual information as in expression (2). This metric is calculated with the probability of words w_i and w_j co-occurring in a document, $P(w_i, w_j)$ and the marginal probability of each word $P(w_i)$.

$$f(w_i,w_j) = \frac{log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-log P(w_i,w_j)} \tag{2}$$

Given that a topic conveys the words in the vocabulary n, Thus, the normalized pointwise mutual information of the kth topic, $TC(k)$, is assessed extending (2) averaging the relatedness of the pair of words involved in the topic as in (3), involving the upper triangular matrix of relatedness between words. Note that, in this case the distribution of topic k are involved in (2) leading to notation $f_k(w_i, w_j)$. For the average, note that with topics involving n words, the number of pairs of words involved are just $\frac{n(n-1)}{2}$.

$$TC(k) = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_k(w_i,w_j) \tag{3}$$

Finally, the normalized pointwise mutual information of the topic model, TC, is assessed as the averaged per-topic coherence attained, as in (4).

$$TC = \frac{1}{K} \sum_{k=1}^{K} TC(k) \tag{4}$$

3.3. Topic to ICD association

Having assessed the coherence of the topics with the idea that each topic should bring semantically related words, our next concern is if the topics inferred are related to the ICDs. That is, each component of the topic-to-document vector enlightens a topic and we wondered if there is a latent relationship between n topics and an ICD. This would mean that the representation itself would be able to discover ICDs.

The topic vector for a given document d conveys the mixture of topics associated to that document, i.e. $\mathbf{v}^d = (v_1^d, \ldots, v_k^d, \ldots, v_K^d)$. $\mathbf{v}^d$ is a vector of size K (the total number of topics), with v_k^d indicating the probability of the topic k for dth document. Therefore, the location in the topic-space of the topics with respect to label L_i, is computed as the prototype topic-to-document vector of the documents including the label L_i, i.e. $\mathbf{c}_i$. This is shown in (5) with $\delta(L_i, d) = 1$ if the ICD L_i was present in EHR d and $\delta(L_i, d) = 0$ otherwise. If the label-set show independent prototypes, then, the prototypes themselves would be able to distinguish the labels. We will show the results attained graphically in Section 4.2.

$$\mathbf{c}_i = \frac{\sum_{d=1}^{D} \delta(L_i,d)\mathbf{v}^d}{\sum_{d=1}^{D} \delta(L_i,d)} \tag{5}$$

4. Experimental results

In this section, we present the experimental results of the Spanish and English datasets. In Sections 4.1 and 4.2, the Topic coherence and Topic to ICD association are explored, respectively. Then, we assess the capacity of the LDA representation to work as features with a plan classifier in Section 4.3 and discuss the results in Section 4.4

The experimental results were carried out using the four ICDs presented in Table 1. The number of topics, K is the dimension of the vectors employed to represent the document. For LDA, the K refers merely to the total number of topics that have to be inferred. However, for PLDA, as it is supervised, it refers to the number of topics dedicated for each label.

This experimental setup aims to compare the obtained representations in Spanish and English, focusing on Spanish, the language of our in-house dataset, and using the MIMIC, in English, as reference. In both cases, the number of topics is set to 48, so the models use the same number of latent variables. Thus, in LDA, we set $K = 48$, and in PLDA $K = 12$, as there are four labels. To determine the number of topics, we used a parameter sweeping approach in order to optimize topic coherence in both for each representation yielding that 48 topics is the optimal value. The PLDA and LDA topic models were generated using the tomotopy library (Fenstermacher, 2020).

Table 4
Topic coherence. L1, L2, L3 and L4 are representative of different specialties: cardiology, neurology and endocrinology.

	Spanish	English
LDA	0.196	0.290
PLDA	0.190	0.193

(a) Global topic-coherence

PLDA	Spanish	English
L1	0.21	0.28
L2	0.23	0.25
L3	0.27	0.24
L4	0.21	0.19

(b) Topic coherence per label in PLDA representation for the examples in Table 1

4.1. Topic coherence

The topic coherence of each representation, LDA and PLDA, is shown in Table 4a. Topic coherence was assessed, as in expression (4). To this end, the most salient n=10 words were taken into account rendering the majority of the total probability mass in the antecedents (Röder, Both, & Hinneburg, 2015; Stevens et al., 2012). The results show that LDA gets better coherence in both datasets. In MIMIC dataset the difference between the coherence in LDA and PLDA is considerable, it seems as if none supervision resulted in beneficial.

As PLDA generates topics per label it is possible to calculate the Topic Coherence for each label. This approach is helpful to see if the topic coherence for each label is stable. Following the labels set as an example in Table 1, in Table 4b is shown that the coherence is similar in those labels of both datasets, being L4 the label with least coherence. Note, as well, that label L4 was the least represented in both data sets (see Table 2). LDA representation is unsupervised so it does not need the labels to generate the representation and therefore the representation itself does not provide a one-to-one relationship between the inferred topics and the labels, so topics per label cannot be calculated just employing the representation as in PLDA.

4.2. Topic to ICD association

In order to assess the correlation between automatically extracted topics and ICDs, the prototypical vectors, $\mathbf{c}_i$, are graphically depicted as a heat map in Fig. 3. Each vector is in a row and its' components are given. This figure shows the topics on the $x-axis$ and the labels or human keywords in the $y-axis$. The darker the cell the higher the presence of the topic.

For example, note that $\mathbf{c}_4$, the prototype vector associated with L_4 brings a salient presence of topics 0 to 11 and nearly no presence of the remaining topics. This pattern, the presence of some topics with the absence of some others, is a characteristic of the ICD with L_4. That is, each component of the topic-to-document vector enlightens a topic and we wondered if each topic is, as well, related with an ICD. This would mean that the representation itself would be able to discover ICDs. This fact is also observed on the other labels. Fig. 3 shows, intuitively, that the topic vectors are prototypical for the labels. Intuitively, the saliency of several sets of topics suffices to connect the EHR with an ICD. Furthermore, multiple labels can be assigned to a document enabling multi-label classification (bear in mind that each EHR tends to convey more than one ICD). These results are extensible for both Spanish and English.

This is a study limited to a set of labels, however, the motivation was to find these types of patterns intuitively and we would not have been able to with a larger label-set. Moreover, the aim was to explore the behaviour in two different though comparable tasks to get solid impressions while being able to cope with an intuitive error analysis.

Table 5
Configurations considered for the multi-layer perceptron.

Configuration	Neurons per layer
C1	(30, 10, 4)
C2	(48, 45, 40, 38, 30, 28, 4)
C3	(1000, 100, 4)

4.3. Downstream-task assessment

So far we have assessed the performance of the topic models themselves to convey coherent meaning and also assessed the ability of the topics to disclose ICDs. In this section topics have been tested as predictive features for supervised classifiers. This task has been performed using a simple supervised classification approach to put the attention on the features instead of complex inference approaches. In this case, both the multi-layer perceptron (Hinton, 1990) and multi-Label k-Nearest Neighbour (ML-k-NN) (Zhang & Zhou, 2007) were considered. Preliminary experiments showed that the perceptron behaved slightly better than ML-k-NN. Thus, for the sake of simplicity, we shall skip the results attained with the ML-k-NN.

The multi-layer perceptron was trained using the configurations in Table 5. The notation used to define the configurations describes the quantity of neurons used in each layer. This configurations differ from each other in therms of the number of layers and quantity of neurons per layers. C1 only has three layers with a small amount of neurons in each. C2 is a 7 layer configuration with small number of neurons per layer. C3 is the configuration that attained the best results and just has two large layers. Thus, we limit the results to the classifier with configuration C3.

Table 6 shows the performance attained by the aforementioned multilayer perceptron employing either LDA or PLDA as predictive features. These results were obtained using both datasets reduced to 4 labels as in Table 1. Precision, Recall and F-score metrics are given by label and, below, the weighted average is given.

The lowest F-score is attained for label L4, the label with the smallest coherence. In Osa dataset, results are slightly better in both configurations compared with MIMIC-III, this can be due to the fact that Osa dataset contains more documents to train the perceptron.

In Spanish and English datasets the same task has been performed using the corpus with 104 labels in Spanish and 50 labels in English. In this case, in the LDA model the number of topics has been established in K=416 and K=200 respectively, which is four times the amount of labels. In the case of PLDA, 4 topics per label were employed, enabling to use the same number of latent variables by both LDA and PLDA models. The results obtained with these corpora show an important drop in the Weighted average F-score. On the one hand, English results show a 70% drop in both representations, this can be due to the fact that the classifier struggle to learn from a model with a higher amount of labels. On the other hand, Osa scores show 48% decline in LDA and 60% in PLDA. The major drop in PLDA may be owing to the higher amount of labels, proving that an unsupervised model behaves better in corpora with a large amount of labels.

To assess the stability of the results while following a train/test evaluation scheme, we initially repeated a representative experiment ten times and evaluated the variance and standard deviation of the metrics. The standard deviation was 0.019 for the F-Score, therefore, the results are stable.

4.4. Discussion

In other domains, (Gangavarapu et al., 2020) employed LDA and deep neuronal networks in order to predict ICD-9 codes of clinical data, and Dermouche, et al. (2016) focused on death certificate classification with documents in French, they prove that LDA can be very intuitive as it allows for a comprehensible representation.

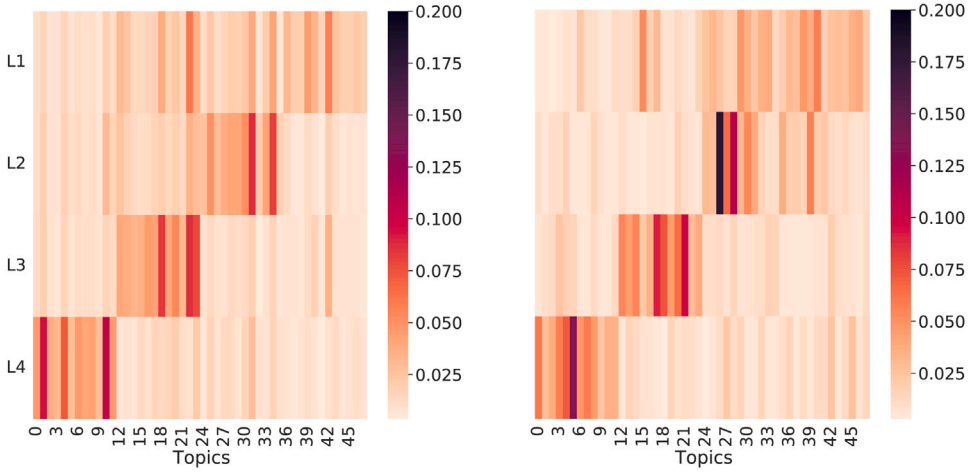


Fig. 3. Topic-to-ICD heatmap showing the presence of each Diagnostic Term or ICD (ordinate) in each topic (abscissa). L1, L2, L3 and L4 are representative of different specialties: cardiology, neurology and endocrinology.

Table 6
Prediction performance for MIMIC-III (English) and Osa (Spanish) datasets. L1, L2, L3 and L4 are representative of different specialties: cardiology, neurology and endocrinology.

	LDA			PLDA		
	Precision	Recall	F-score	Precision	Recall	F-score
L1	0.35	0.12	0.18	0.28	0.97	0.43
L2	0.43	0.09	0.15	0.35	0.75	0.47
L3	0.65	0.88	0.75	0.64	0.68	0.66
L4	0.39	0.27	0.32	0.20	0.01	0.01
W Avg.	0.50	0.45	0.43	0.43	0.62	0.45

(a) MIMIC dataset

	LDA			PLDA		
	Precision	Recall	F-score	Precision	Recall	F-score
L1	0.46	0.43	0.44	0.50	0.43	0.46
L2	0.61	0.63	0.63	0.70	0.64	0.67
L3	0.76	0.73	0.74	0.76	0.82	0.79
L4	0.35	0.32	0.33	0.52	0.30	0.38
W Avg.	0.63	0.61	0.62	0.67	0.65	0.66

(b) Osa dataset

The authors who focused primarily on the supervised classification, such as Miranda-Escalada et al. (2020) in Spanish, and Névéol, Robert, et al. (2018) or Vaswani, et al. (2017) in English, do not pay attention to the representation and lack the intuition to know whether or not they are getting a representation associated with the ICD. Thanks to the measures of association vector in Eq. (5) we understand how much it is associated each vector with the labels.

We prove that even though LDA representations can discover latent topics, there is no guarantee that the topics convey human intuition of keywords and, hence, meet expectations. In the case of EHR representation, clinicians could expect as keywords the ICDs. In contrast to LDA, PLDA promotes topic-to-ICD semantic consistency. Comparing LDA and PLDA, we found that PLDA is more coherent and is well correlated with the ICD.

Attending to their predictive performance (Table 6), PLDA shows better F-score results than LDA in both datasets. In both cases the smallest F-score is obtained when trying to predict the L4 label. We hypothesize that it is related to the topic coherence, as the L4 was the label which had the least coherent topics related. Nevertheless, PLDA fails in this aspect when dealing with corpora with a larger label-set. In this case LDA seems to work better. Our hypothesis is that there is not enough text (i.e., not enough EHRs) related to each label for the Supervised model to extract meaningful associations.

Table 7 depicts the topic-to-word vector (the mixture of words) restricted to the 10 most common words of the topic $L1$, diabetes mellitus in MIMIC and Osa corpora. The weight column of each word corresponds to its probability in the topic, thus the smallest weights are the least representative. The topic in both tasks, MIMIC (English) and Osa (Spanish), contains words directly related with diabetes: *diabet*, *sugar* and *insulin* in MIMIC-III and *glucos*, *globulin* and *insulin* in Osa. The query of the ICD codes related to "diabetes mellitus" show a significant overlap among the words in the ICD terms and standard descriptions and the highest weighted words from PLDA.

With topic models representing a document as a mixture of topics and a topic as a mixture of words (Table 7), then, note the parallelism with the attention mechanisms (Vaswani, et al., 2017) that account the relevance of each to motivate the prediction of a label. In this sense, topic models can also point out the most salient words. In previous papers LDA was used to predict interpretable topics in large documents (Gao, et al., 2020).

5. Conclusions

This work deals with clinical text mining in Spanish. Building resources for languages other than English involves particular challenges (Névéol, Dalianis, et al., 2018). In particular, we tackled the EHR

Table 7
Topics inferred by PLDA related with label L1, diabetes mellitus W/o mention complication.

MIMIC		Osa	
Word	Weight	Word	Weight
Insulin	0.87	Colesterol	1.10
Diabet	0.56	Globulin	0.76
Sugar	0.48	Glucos	0.56
Nausea	0.35	Leucocito	0.52
Lantus	0.38	Creatinin	0.52
Humalog	0.33	Hematie	0.48
Hyperglycemia	0.30	Linfocit	0.47
Ketoacidosi	0.29	Hemoglobin	0.40
Glargin	0.27	Insulin	0.35

multi-label classification problem according to the ICD in Spanish and compared the methods employed with English. The aim is to classify the EHRs utilizing the text from the records. In this task, the representation is crucial to convey valuable features that render the semantics. In this preliminary work, we focus on the representation of the text. In particular, we pay attention to interpretable representations to aid the clinicians involved in the loop.

In Spanish text mining, creating accurate word-embeddings for the clinical domain might be difficult due to the lack of specific resources. This issue led us to discard approaches based on embeddings. PLDA and LDA topic models have been applied on two different datasets of discharge summaries to get their representation. We conclude that PLDA deduces more coherent topics than LDA. We have also applied the representation as input for a downstream task model, i.e., a multi-label DNN-based classifier. PLDA shows better behaviour in terms of the F-score in the two datasets. However, when the amount of labels increases, the coherence drops off a 27% and the F-score decreases by 60%. Both LDA and PLDA make the most of small corpora and should not be discarded in domains or languages with few corpora or linguistic resources.

Our next goal is to carry out the experiments focusing on the entire set of labels and, eventually, carry out a quantitative comparison with embedded topic models. We have the impression that multi-lingual topic models can enhance tasks with few resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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 by the Basque Government, Spain (IXA IT-1343-19, Predoctoral Grant PRE-2019-1-0158). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

The authors would like to thank the anonymous reviewers for carefully reading the manuscript and providing constructive suggestions that helped to improve our work.

Appendix. Research steps

A.1. Document pre-processing

Following the background trend, the datasets have been pre-processed using *stopwords* followed by a *lemmatizer* and *stemmer*.

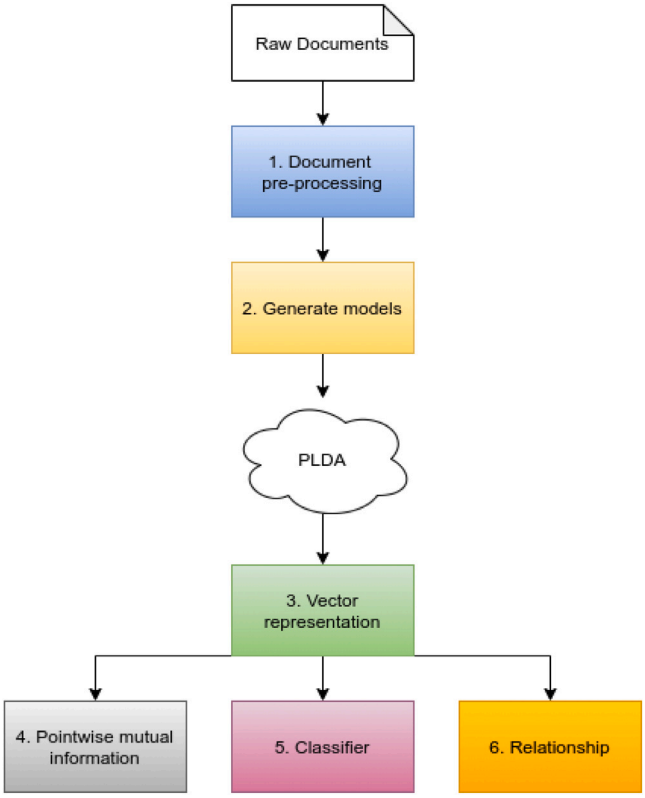


Fig. A.4. Flow diagram of the research.

- **Stopwords**: Very common words that have no relevant meaning, prepositions and determiners, for instance, (*a, to, is, on, the* ...) are considered *Stopwords*. The Gensim (Řehůřek & Sojka, 2010) library has been used to remove these words from the text.
- **Stemmer:** This technique consists of removing the endings of the words in order to try to obtain only their roots. For example for the word *stairs* the Stemmer would return *stair*. This has been implemented using the SnowballStemmer from the nltk library (Bird, Klein, & Loper, 2009).
- **Lemmatizer:** Is a technique similar to Stemmer, but it tries to perform a morphological analysis of the words and return the base of a word, i.e. its lemma. Normally this technique is used together with Stemmer to obtain better results. Unlike the Stemmer, the Lemmatizer would return the word 'run' having obtained the word 'ran'. It has been implemented using the WordNetLemmatizer from the nltk library (Bird et al., 2009).

A.2. Flow diagram and libraries

Fig. A.4 shows the outline of the steps taken in the research employing PLDA representation. The outline using LDA is similar. In each step we have used the following libraries:

1. Document pre-processing: As explained before, in this step we have used Gensim (Řehůřek & Sojka, 2010) and nltk (Bird et al., 2009) libraries.
2. Model generation: the models needed to generate the topic vectors from the documents have been generated using the Tomotopy library (Fenstermacher, 2020).
3. Vector representation: Using the models and Tomotopy library the vector representation of the text has been generated.

4. Pointwise mutual information: The normalized pointwise mutual information of the vectors have been computed following Eq. (3) in Section 3.2.
5. Classifier: In order to assess the downstream task, both ML-k-NN and multi-label perceptron have been used. We have used the implementation available in scikit-multilearn (Pedregosa, et al., 2011) python library.

References

Abdulaziz, W., M. Ameen, M., & Ahmed, B. (2019). An overview of bag of words;importance, implementation, applications, and challenges. (pp. 200–204).

AlShuweihi, M., Salloum, S. A., & Shaalan, K. (2021). Biomedical corpora and natural language processing on clinical text in languages other than english: A systematic review. In *Recent advances in intelligent systems and smart applications* (pp. 491–509). Springer.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. Beijing: O'Reilly.

Blanco, A., Pérez, A., & Casillas, A. (2020). Extreme multi-label ICD classification: Sensitivity to hospital service and time. *IEEE Access*, *8*, 183534–183545.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–480.

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing topic models. In *ICWSM*.

Chen, Q., Peng, Y., & Lu, Z. (2019). BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE international conference on healthcare informatics* (pp. 1–5). IEEE.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

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* (pp. 61–68).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, Volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., & Grune, B. (2019). Overview of the CLEF ehealth 2019 multilingual information extraction. CEUR-WS.

Fenstermacher, D. (2020). Tomotopy: Gibbs-sampling based topic model python library extension of tomoto (topic modeling tool). Zenodo.

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

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., et al. (2020). The pile: An 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., et al. (2020). Overview of the CLEF eHealth evaluation lab 2020. In *International conference of the cross-language evaluation forum for European languages* (pp. 255–271). Springer.

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine learning* (pp. 555–610). Elsevier.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1–2), 177–196.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 1–9.

Kalyan, K. S., & Sangeetha, S. (2020). Secnlp: A survey of embeddings in clinical natural language processing. *Journal of Biomedical Informatics*, *101*, Article 103323.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., & Krallinger, M. (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*.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. CoRR, arXiv:1802.05695.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, *9*(1), 12.

Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., et al. (2018). CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in french, hungarian and Italian. In *CLEF*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 457–465).

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM international conference on web search and data mining* (pp. 399–408).

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning*, *88*(1–2), 157–208.

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 145–158). Springer.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952–961).

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 I2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Voorhees, E. M., & Hersh, W. R. (2012). Overview of the TREC 2012 medical records track. In *TREC*.

Wang, Y., Bai, H., Stanton, M., Chen, W.-Y., & Chang, E. Y. (2009). Plda: Parallel latent dirichlet allocation for large-scale applications. In *International conference on algorithmic applications in management* (pp. 301–314). Springer.

Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048.