lebena_2022_preliminary_exploration_of_topic_modelling_representations_for_electronic_health_records_coding_according_to_the_international_classification_of_diseases_in_spanish

Year

2022

Author(s)

Nuria Lebena and Alberto Blanco and Alicia Perez and Arantza Casillas

Title

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

Venue

Expert Systems with Applications

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Result of supervised topic modeling (PLDA)

Topic labeling parameters

Label generation

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. 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.

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

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

Motivation

"the saliency of several sets of topics suffices to connect the EHR (the document) with an ICD (the topic-label). 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)."

Topic modeling

PLDA

Topic modeling parameters

Nr of topics dedicated to each label: 12
(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.)

Nr. of topics

48 (12 x 4 labels)

Label

One of four ICD codes (L1-L4): Diabetes mellitus W/o mention complication, Congestive heart failure, Congestive heart failure, Acute kidney failure
(Note that labels refer to Diagnostic Term (DT), and each DT is encoded using an ICD)

Label selection

Topic to ICD association

The topic vector for a given document d conveys the mixture of topics associated to that document:

$$\mathbf{v}^d = (v_1^d, \ldots, v_k^d, \ldots, v_K^d).\ \mathbf{v}^d$$

is a vector of size K (the total number of topics), with

v_k^d

indicating the probability of the topic k for d-th 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

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

with $\delta(L\_i, d) = 1$ if the ICD L_i was present in EHR d and $\delta(L\_i,d) = 0$ otherwise.

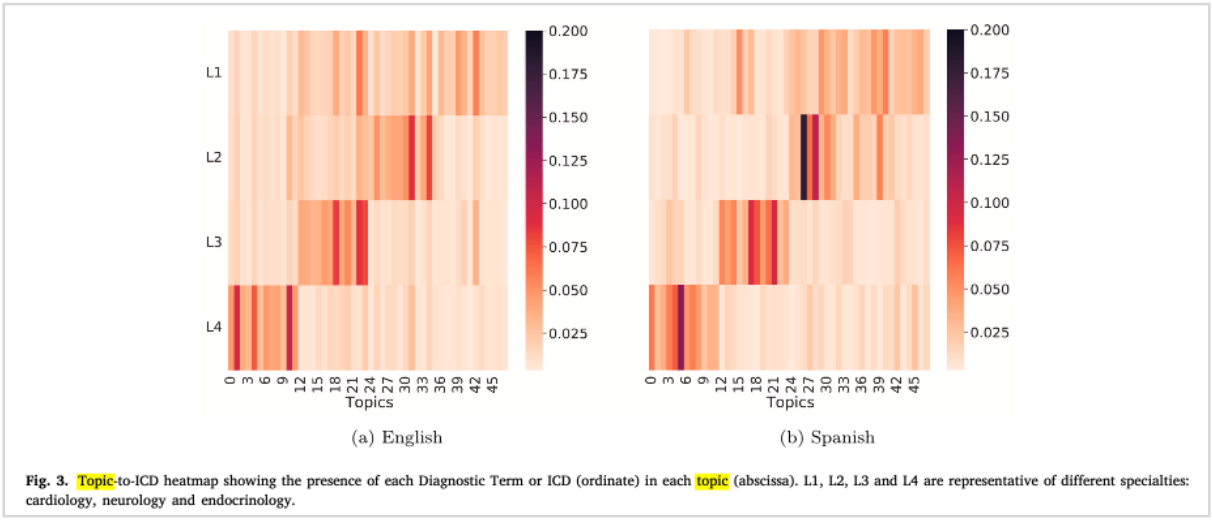

Fig. 3. Topic-to-ICD heatmap showing the presence of each Diagnostic Term or ICD (ordinate) in each topic (abscissa). L1, L2, L3 and L4 are representative of different specialties: cardiology, neurology and endocrinology.

Label quality evaluation

\

Assessors

\

Domain

Paper: Health
Dataset: Electronic Health Records

Problem statement

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.

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.

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.

Corpus

Dataset 1
Origin: Public health system (Spain)
Nr. of documents: 22611
Details:

- 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.

Dataset 2
Origin:
Nr. of documents: 11519
Details:

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

Document

Pre-processing

```
@article{lebena_2022_preliminary_exploration_of_topic_modelling_representations
_for_electronic_health_records_coding_according_to_the_international_classifica
tion_of_diseases_in_spanish,
  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.},
  author = {Nuria Lebe{\~n}a and Alberto Blanco and Alicia P{\'e}rez and Arantza Casillas},
  date-added = {2023-03-22 16:55:13 +0100},
  date-modified = {2023-03-22 16:55:13 +0100},
  doi = {https://doi.org/10.1016/j.eswa.2022.117303},
  issn = {0957-4174},
  journal = {Expert Systems with Applications},
  keywords = {Multi-label classification, Document classification, Electronic Health Records, ICD classification, Topic models, Partially labelled dirichlet allocation},
  pages = {117303},
  title = {Preliminary exploration of topic modelling representations for Electronic Health Records coding according to the International Classification of Diseases in Spanish},
  url = {https://www.sciencedirect.com/science/article/pii/S0957417422006662},
  volume = {204},
  year = {2022}}

#Thesis/Papers/Initial