Extraction and Representation of Clinical Terms from Portuguese Clinical Text

Veniamin Craciun
Instituto Superior Técnico, Universidade de Lisboa, Portugal
veniamin.craciun@tecnico.ulisboa.pt

November 2019

Abstract

The adoption of electronic health records (EHRs) in clinics and hospitals made possible to aid healthcare professionals on a daily basis allowing to take decisions faster and better. However, EHRs often consist of large quantities of unstructured data, such as medical notes in free-text, which given their nature, contain several medical conditions, symptoms or clinical terms which could or could not be related in the same medical note, hence leading to ambiguity as to which are the most important terms and how these relate to each other. This work proposes an architecture capable of processing text contained in clinical documents, extracting the key-concepts contained in them. This is accomplished through application of dimensionality reduction techniques, that transform the original high-dimensional data to a latent representation extracting only the most important features, such as non-negative matrix factorization and deep neural networks. After extraction and latent representation of key-concepts, is possible to analyze which attributes are the most similar and how these relate to each other.

Keywords: clinical notes, clinical terms extraction, similarity, dimensionality reduction

1. Introduction

The introduction of Electronic Health Record (EHR) in clinics and hospitals, made possible to store and to aggregate clinical information from patients, containing structured and unstructured data. An example of unstructured data is the textual description, written by a healthcare professional, of a patient reason for encounter (RFE) or complaint, i.e., a concise statement describing the symptom or problem. Challenges of working with of unstructured data is high-dimensionality, temporality which refers to the sequential nature of clinical events, irregularity which refers to high variability, bias including systematic errors in the medical data, and mixed data types and missing data [Sadati et al., 2019]. Moreover, other challenges are related to those found in other Natural Language Processing (NLP) problems when dealing with human generated free-text, namely abbreviations, orthographic errors and ambiguities. In the NLP field these problems have been studied and can be tackled, to some degree, by employing word embeddings, which have been used to extract medical terms [Vine et al., 2015], to encode medical terms to compute similarities between patients [Choi et al., 2016, Gencoglu, 2019, Zhang et al., 2019, Zhu et al., 2016] or to find co-occurrences of health conditions [Bhattacharya et al., 2016, Gefen et al., 2018].

Taking inspiration from previous works, we propose to create an architecture that processes textual information contained in medical documents, extracting the key-concepts through AutoPhrase, a key-extraction framework, and these key-concepts and original documents are transformed to a latent representation through dimensionality reduction techniques. After this transformation, the amount of original data contained in the medical documents is condensed, hence removing any irrelevant or noisy information. Furthermore, the new latent representation allows to compare how the extracted keywords relate to each other. In the medical domain, the keywords extracted could be either medical conditions, symptoms or any other clinical term used by healthcare professionals.

Moreover, we study the performance of multiple dimensionality reduction techniques, based on Non-negative Matrix Factorization (NMF) and deep neural networks, on the ability to create latent representation and to capture the similarity between the keywords extracted, assessing if key-concepts given by healthcare professionals follow the hierarchical structure of two classification systems, namely International Classification of Diseases 10th revision (ICD-10)[1] and 2nd version of International Classification of Primary Care (ICPC-2)[2]. Both classification systems are composed of multiple chapters where each chapter usually ag-

[1] `https://www.who.int/classifications/icd/en/`

[2] `https://www.who.int/classifications/icd/adaptations/icpc2/en/`

gregates related conditions or diseases. In order to assess their hierarchical structure, we propose to compare if keywords or codes from the same chapter are more similar than those from different chapters. If that is the case, it indicates that the keywords and codes assigned by healthcare professionals follow the hierarchical structure of the classification system.

This work aims at answering the following research questions: RQ-I) *Which technique best preserves the hierarchical structure of ICD-10 and ICPC-2?*, i.e., if codes and key-phrases from the same chapter show higher similarity than those from different chapters. RQ-II) *Which dimensionality reduction technique performs the best from a topic modelling and clustering perspective?*

The remaining of this document has the following structure. Section 2 introduces the theoretical concepts of the methods developed in this work and the literature overview of similar work developed by other teams, from which the current work takes inspiration. Section 3 details the architecture implemented in this work. Section 4 presents the statistical characterization of datasets and evaluation metrics used, with the respective achieved results discussed in Section 5. Closing with the main contributions and limitations of the present work and other ideas to consider for future work in Section 6.

2. Related Work

With the introduction of Electronic Health Record (EHR), large quantities of data become available for analysis. However, when dealing with such quantities it is inevitable that the information is not structured containing bias and unbalanced. To solve these issues, multiple approaches have been proposed in the bioinformatics and machine learning literature which aim at encoding the meaning contained on documents from the clinical domain by learning the underlying representation of the data. This section describes existing methods proposed in the literature to overcome these issues leveraged by matrix factorizations and deep learning techniques.

2.1. Non-Negative Matrix Factorization

One of the most known dimensionality reduction technique through matrix decomposition is Non-negative Matrix Factorization (NMF), introduced by Lee and Seung [2000], and has been widely applied in a plethora of fields, ranging from astronomy, audio signal processing, bioinformatics, computer vision and text mining. To better understand it, let us consider a data matrix $C \in \mathbb{R}^{v \times d}$ with v dimensions and d data-points which has only non-negative elements. In text mining applications, C corresponds to a document-term matrix constructed with the weights of v terms (e.g., words and/or key-phrases, typically weighted according to a heuristic such as TF-IDF) from a set of d documents. If we define two matrices, also with only non-negative elements, respectively $W \in \mathbb{R}^{v \times r}$ and $H \in \mathbb{R}^{r \times d}$, then the NMF technique can reduce the dimensionality of C through the approximation:

$$C \approx W \cdot H, \text{ generally with } r < \min(v, d) \tag{1}$$

where the columns of W make up the new basis directions of the dimensions we are projecting onto and each column of H represents the coefficients of each data point in this new subspace. In text mining applications, it is useful to think of each feature (i.e, each column vector) in the features matrix W as a document archetype, comprising a set of tokens where each tokens's cell value defines its rank in the feature (i.e., the higher a tokens's cell value, the higher its rank in the feature). A column in the coefficients matrix H represents an original document with a cell value defining the document's rank for a feature. Through NMF technique, we can reconstruct a document (i.e., a column vector) from the input matrix C by a linear combination of the features (i.e., column vectors in W) where each feature is weighted by the feature's cell value from the document's column in H. The factorization is usually sought after minimization of the reconstruction between original matrix C and new matrices W and H as an optimization problem, as follows:

$$\min_{W,H} D(C \| W \cdot H), \text{subject to } W \geq 0, H \geq 0 \tag{2}$$

where $D(\cdot)$ is known as the cost function, which is defined by:

$$D(C \| W \cdot H) = \sum_{ij} d(C_{ij} \| (W \cdot H)_{ij}) \tag{3}$$

having $d(x \| y)$ as the scalar cost function. In the literature, several scalar cost function have been proposed, most commonly used are the original proposals made by Lee and Seung [2000], which are based on Frobenious norm (i.e., extension of Euclidean norm to matrices), defined as:

$$d_{FR}(x \| y) = \frac{1}{2}(x - y)^2 \tag{4}$$

and the generalized Kullback-Leibler divergence, which takes the following form:

$$d_{KL}(x \| y) = x \log \frac{x}{y} + (y - x) \tag{5}$$

Both cost functions are positive and take value zero if and only if $x = y$, i.e., when the reconstruction of the original matrix is perfect. In our experiments, we will study the results produced by each cost function described previously.

Several extensions have been proposed to the standard NMF algorithm. For instance, SNMF

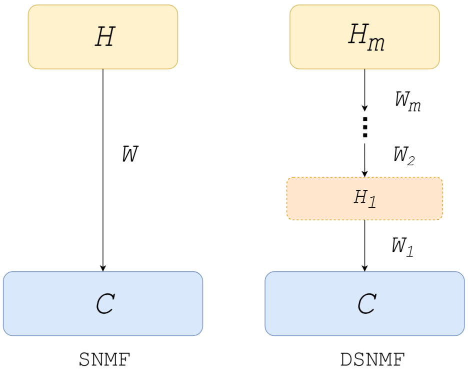

Figure 1: Comparison between SNMF and DSNMF, in which the matrix C is factorized in m steps.

[Ding et al., 2010] is a matrix factorization technique that learns a low-dimensional representation of a dataset that lends itself to a clustering interpretation. The SNMF differs from NMF because the matrix W has no restriction on the values that it can take, i.e., $W \in \mathbb{R}_{\pm}^{v\times r}$ while the restriction on H maintains to non-negative values, i.e., $H \in \mathbb{R}_{+}^{r\times d}$. The authors argue that SNMF helps to close the bridge between NMF and k-means clustering. In this case the matrix W serves as the cluster centroids while H can be viewed as the cluster indicators for each data point.

Trigeorgis et al. [2014] proposed Deep Semi-NMF (DSNMF) arguing that the SNMF algorithm does not capture lower-level hidden attributes in a dataset that contains complex hierarchical information. Hence, by factorizing the matrix H in multiple steps, one is able to find these lower-level hidden attributes, in terms of clustering interpretation. Figure 1 allows us to better understand the main differences between the SNMF model and DSNMF. The former is a simple linear transformation of the initial input space, whereas the latter uses deep neural networks that learn the hierarchy of hidden representations. For instance, by employing the latter technique, one is able to find low-level features from a human face (e.g., facial expression or pose). This can be perceived as successively factorizing a representation matrix C into m steps, such that:

$$C^{\pm} \approx W_1^{\pm} \cdot W_2^{\pm} \dots W_m^{\pm} \cdot H_m^{+} \tag{6}$$

The previous equation can also be understood as if the SNMF has been applied successively over the matrix H at step $m-1$, resulting in matrices W and H at step m, as shown in the following equations:

$$\begin{gathered} H_{m-1}^{+} \approx W_m^{\pm} \cdot H_m^{+} \\ \vdots \\ H_1^{+} \approx W_2^{\pm} \cdot H_2^{+} \end{gathered} \tag{7}$$

The main difference of this model was the introduction of a non-linear function $g(\cdot)$ to approximate the original input matrix C using H. Hence, H_m is approximated, as follows:

$$H_m \approx g(W_{m+1} \cdot H_{m+1}) \tag{8}$$

The authors proposed to use a modified version of the Frobenius norm (Equation 4) as the optimization problem, minimizing $D(C\|W \cdot H)$ which in this particular case is defined as:

$$\sum_{ij} d_{FR}(C_{ij}\|(W_1 \cdot g(W_2 \cdot g(\dots g(W_m \cdot H_m)_{ij}))) \tag{9}$$

Leveraging this method, the authors showed that DSNMF is able to cluster specific features of human faces at different levels. For instance, H_1 holds information regarding the pose of a face, whereas H_2 contains features regarding the facial expression, and H_3 represents the identity of a person. They also show, thorough examples, that their approach is able to outperform not only SNMF but also other NMF variants at learning low-dimensional representations for clustering purposes.

2.2. Autoencoder

Another unsupervised dimensionality reduction technique is the Autoencoder (AE), proposed by Hinton and Salakhutdinov [2006] as a non-linear generalization of PCA more capable at extraction the relevant features from data, outperforming PCA in clustering tasks. The autoencoder is composed of three main parts, namely the encoder, latent representation and the decoder. The encoder is used to transform highly dimensional data into a smaller latent space, while the decoder is used to transform back the learned features from the latent space back to obtain a reconstruction of the original data. Both encoder and decoder can contain one or several layers, forming deep neural networks.

To better understand this technique, let us sample a data-point $x \in \mathbb{R}^{v\times 1}$ from C and feed it to the encoder network, then latent representation is produced by $h = g\left(W_1 \cdot x + b\right)$, where g$(\cdot)$ is an element-wise linear or non-linear function with non-negative outputs, where $b \in \mathbb{R}^{r\times 1}$ is the bias term, while $h \in \mathbb{R}^{r\times 1}$ is the representation in the latent space, and having $W_1 \in \mathbb{R}^{v\times r}$ as the weights of the first layer of the encoder network. It is also possible to add additional layers to make a deeper network with multiple hidden layers before arriving at the constriction layer which produces the h output. The final set of weights must be kept non-negative, which can be achieved with an identity activation function $f(\cdot)$, such that $\hat{x} = f(W_f \cdot h)$, where $W_f \in \mathbb{R}^{m \times r}$. The final weights of the autoencoder can be interpreted as the dimensions of the new subspace, with the elements of h as the coefficients in that subspace. The encoder and decoder networks are trained so that $x \approx \hat{x}$, through a loss function such as the mean square error (MSE), described by:

$$\mathcal{L}_{AE} = \frac{1}{2} \sum_{i=1}^{n} (x^{(i)} - f(g(x^{(i)})))^2 \qquad (10)$$

which minimizes the reconstruction error between the input and the output.

2.3. Key-phrase Extraction

One of the simplest and fastest key-phrase extraction baseline is TF-IDF, where each key-phrase has assigned a score and ranked according to the its occurrence in a document and overall occurrence in the dataset. However, more recently several approaches have been proposed that take advantage of the co-occurrence statistics from external sources as is the case of AutoPhrase [Liu et al., 2015, Shang et al., 2018]. The novelty of this approach is that it allows to extract relevant information from documents given a knowledge base (i.e., terminological resources such as Wikipedia) available on the targeted language. Moreover, AutoPhrase does not require human intervention to select quality phrases, because these are automatically selected from the provided knowledge base.

In order to select quality phrases, the first step consists of building a candidate set of potential phrases based on their n-gram frequency, by imposing a threshold for frequency or by choosing a value for n. After the candidate set has been created, AutoPhrase uses the knowledge base to create two pools of phrases, namely a positive pool and a noisy negative pool based on the quality of a given phrase. The quality estimation is done taking into account the following four metrics:

- **Popularity:** quality phrases should occur with sufficient frequency in the given document collection, e.g., currently the word *database* is more popular than the original version *data base.*

- **Concordance:** the collocation of tokens present in quality phrases occurs with significantly higher probability than expected due to chance, e.g., the phrase *strong tea* is considered more concordant than *powerful tea* because the former is used more frequently.

- **Informativeness:** a phrase is informative if it is indicative of a specific topic or concept, e.g., *this paper* is a popular and concordant phrase, but does not add any additional information in a research publication corpus.

- **Completeness:** long frequent phrases, and subsequences from those phrases, may both satisfy the three criteria above. A phrase is deemed complete when it can be interpreted as a complete semantic unit in some given document context. Note that a phrase and a subphrase contained within it may both be deemed complete, depending on the contexts in which they appear. For example, the phrases *relational database system*, *relational database* and *database system* can all be valid in certain contexts.

We will use AutoPhrase into our implementation to extract the most relevant key-phrases for each document in the dataset, as is described in Section 3.

2.4. Medical Concepts Representation

The methods described previously have been widely applied to create architectures that process documents from the medical domain. For instance, Choi et al. [2016] created Med2Vec, a multi-layer architecture capable of learning code-level and visit-level representations by employing word embeddings, namely the skip-gram model Mikolov et al. [2013] based on co-occurrences of codes and other demographic information of a patient. Leveraging this approach it creates a final representation as a non-negative matrix that allows to find relationships between patients or codes. The authors argue that this representation is highly interpretable, which is crucial factor in healthcare applications, where understanding the obtained results is very important, and the performance of the model can be compromised.

In this line of work, other teams focused focused on identifying co-occurrences of health conditions from clinical documents using topic modelling approaches, through LDA [Bhattacharya et al., 2016] and LSA [Gefen et al., 2018]. Specifically, Bhattacharya et al. [2016] showed that LDA is capable of capturing hidden patterns between complications, e.g., between *diabetes* and *function decrease of kidneys*, among other interesting associations. The topics retrieved were validated by medical literature and proved to relate to real world co-occurrence of health conditions.

3. Methods

Inspired by the techniques aforementioned, we present our proposed approach which is based on matrix factorization techniques, as is the case of

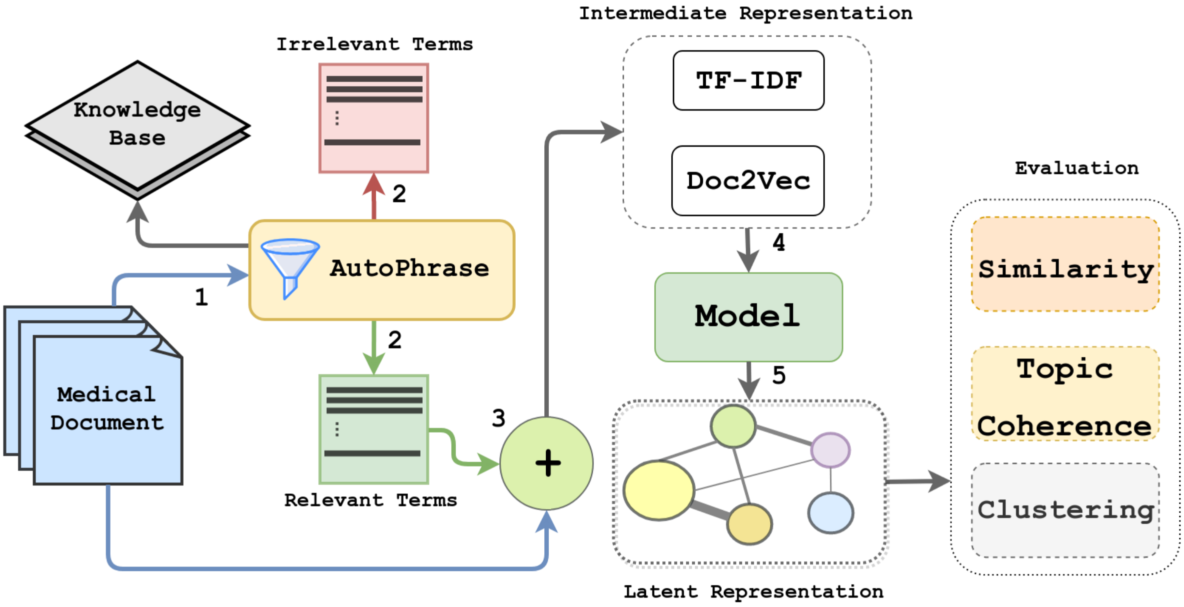

Figure 2: Outline of the architecture, starting from key-phrase extraction with AutoPhrase, transforming the documents to a intermediate vector representation based on either TF-IDF or Doc2Vec, and the final latent dense representation and its final evaluation.

NMF, SNMF, and deep neural networks as is the case of autoencoder and DSNMF previously presented. Our approach is composed of three stages, and a visual representation is depicted in Figure 2. First, a pre-processing stage was in place, it consisted on stop word removal and lowercase transformation, then we feed all documents to AutoPhrase in order to uncover the most relevant key-phrases in each document, providing the ICD-10-CM code descriptions in Portuguese[3] as knowledge base. On the second stage, the original documents and key-phrase extracted are transformed to a intermediary representation (step 3 in Figure 2), leveraging through either TF-IDF terms score or through Doc2Vec method using the distributed memory architecture as proposed by Le and Mikolov [2014].

In the third stage, we apply every method over the intermediary representation (step 4), which yields the final latent representation that is then evaluated based on topic modelling and clustering perspective. In order to compare all methods experimented, the pre-processing and vector space transformation stages are the exactly equal, thus minimizing any variance that could be introduced in these stages that could affect the final result. Follows the implementation details of all methods used in this work.

In our study, we used the open-source implementation of NMF available from the scikit-learn[4] package, which implements the fast multiplicative update solver, proposed by Févotte and Idier [2011]. We also used the a dense variant of the Nonnegative Double Singular Value Decomposition initialization method (NNDSVDa) for matrices W and H, based on two SVD processes in which one approximates the data matrix, and the other approximates positive sections of the resulting partial SVD factors, utilizing an algebraic property of unit rank matrices [Boutsidis and Gallopoulos, 2008]. In our experiments, we used both the Frobenius norm and Kullback-Leibler divergence, which we refer to as FR-NMF and as KL-NMF, respectively. Meanwhile, both SNMF and DSNMF used were open-source[5] implementations using the Tensorflow[6] library based on original proposal by Trigeorgis et al. [2014]. Similarly, we used an adapted version of the open-source autoencoder implementation by the Tensorflow research team[7]. These adaptations were necessary in order to be compatible with pre-existing data flow in architecture pipeline.

We used the same parameters for both DSNMF and AE, both having a two hidden-layer network, with 2000 as the size of the first hid-

[3] https://interop-pt.atlassian.net/wiki/spaces/CTCPT/pages/58884216/Projetos

[4] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

[5] https://github.com/yvettegwy/deep-semi-nmf/

[6] https://www.tensorflow.org

[7] https://github.com/tensorflow/models/tree/master/research/autoencoder

den layer while the second layer (i.e., the latent representation) was parametrized by $\tau \in \{5, 10, 15, 20, 25, 30, 40, 50, 100, 150, 200\}$, we used ReLU$(\cdot)$ as activation function, and the feedforward neural network was trained through the Adam optimizer [Kingma and Ba, 2014] with a learning rate of $\eta = 0.001$ for a 100 epochs and with a tolerance of 0.05 between epochs as stopping condition. We used different values for the latent representation τ in order to compare all methods across different latent dimensionalities, doing so, allows us to verify if these methods display different behaviors when other dimensions are used.

4. Datasets and Experimental Evaluation

This section describes the metrics used to evaluate the architecture on a dataset consisting of free-text clinical descriptions in Portuguese. First, we give a statistical characterisation of the dataset used and then introduce the evaluation methodology employed to measure all methods tested during our experimentation.

4.1. Datasets

The dataset is composed of 206 644 death certificates, hence forward referred to as documents, collected between the years of 2013 and 2016. Each document contains a brief free-text description for the cause of death in Portuguese alongside the respective ICD-10 code for the main cause of death, as well as ICD-10 codes for auxiliary/contributing conditions. Both the free-text description and ICD-10 codes were written and labelled by a healthcare professional. Furthermore, originally the dataset was not labelled accordingly to the ICPC-2 terminology, hence a transformation of the dataset was in place in order to create the mapping between ICD-10 and ICPC-2 systems, resulting in a total of 8 452 documents mapped. During our experimentation, we used all documents mapped to ICPC-2 and the remaining 197 992 documents were filtered to only include ones that were labelled with a ICD-10 code with a frequency higher than 10, from these we sampled 50 000. After removing duplicate documents if the ICD-10 code assigned, free-text description and key-phrases extracted were repeated, yielding a total of 41 219 documents, hence forward referred to as DS_1. We only sampled 50 000, because it was a hardware limitation when running the DSNMF method, as the implementation of this method is computationally heavy during the transformation of the input, and can not be done in several batches as is the case of autoencoder. For testing purposes, we experimented the other methods with all documents but there was not a significant improvement on the results over considering the sub sample consisting of only 50 000 documents.

4.2. Experimental methodology

We decided to evaluate all methods from three different perspectives in order to answer to thse research question RQ-I and RQ-II. One perspective is to verify if the clinical terms and codes used by healthcare professionals align with the hierarchical structure of ICD-10 and ICPC-2 classification systems, this can be achieved by measuring the cosine similarity between collection of pairs of clinical terms and codes. To create these collections, each pair must be composed of two distinct codes, each pair of codes has to be strictly from either classification system, i.e., a pair with one code from ICD-10 and the other ICPC-2 should not occur, neither be compared. To compare key-phrases extracted, we used their respective ICD-10 or ICPC-2 codes to build the pairs. Enforcing these constraints, the following collection of pairs were created, these are:

- Pairs of ICD-10 codes within the same chapter (ICD SC), same chapter and block (ICD SB), from different chapters (ICD DC), from different blocks (ICD DB)
- Pairs of ICPC-2 codes within the same chapter (ICPC SC), same chapter and block (ICPC SB), from different chapter (ICPC DC).
- Pairs of phrases with ICD-10 code assigned within the same chapter (P ICD SC), same chapter and block (P ICD SB), different chapter (P ICD DC), different blocks (P ICD DB).
- Pairs of phrases with ICPC-2 code assigned within the same chapter (P ICPC SC) and different chapter (P ICPC DC).

After creating these collections, we sampled 330 random pairs from each collection and computed the similarity between each pair by using their respective latent representation. We have only considered 330 pairs because it was the maximum number of codes mapped in the P ICPC SC collection. The average cosine similarity results achieved by each method for code-pairs and phrase-pairs collections are presented in Table 2 and Table 3, respectively.

The second perspective is related to topic modelling , i.e., to measure how consistent are the resulting topics created by each methods experimented. This can be achieved by applying topic coherence scores such as CV measure proposed by Röder et al. [2015], which capture the semantic interpretability of the discovered topics, aligning with human evaluations of a topic.

The third perspective is to measure the resulting latent representation in terms of clustering performance. We compare all methods on both on supervised metrics, namely Normalized Mutual Information (NMI) and Homogeneity, and on unsupervised metrics, specifically Davies-Bouldin Index (DBI) and Silhouette coefficient. The supervised metrics allows us to evaluate if there is any correlation between ground truth labels and labels assigned by the clustering algorithm (NMI) and if there are clusters that only contain data points of a single class (Homogeneity). While the unsupervised metrics measure the structure of the clustering, for instance the ratio between intra-cluster and inter-cluster distances with DBI (a value of 0 is best) and measure if there are overlapping clusters with Silhouette coefficient ranging from worse at -1 to best at 1, while a value of 0 is indicative of overlapping clusters, negative values indicate that a sample was assigned to the wrong cluster, as a different cluster is more similar.

To estimate these metrics, we employed k-means clustering algorithm over the τ latent representations, then took the average to obtain an aggregated result. Doing so, we avoid poor initializations which lead to local minima while comparing the methods considering multiple dimesnionalities given by τ. The k-means clustering is parametrized by the number of clusters, in our case we use the number of distinct ICD-10 chapters, and in our particular dataset sample (DS_1) there were a total of 14 chapters.

5. Results

As mentioned previously the performance of each model was evaluated from three perspectives, namely cosine similarity between pairs of extracted terms, measuring the consistency of the topics generated by each model through topic coherence scores, and based on clustering performance. Better performing models will have a higher CV coherence score, as well as being be able to produce higher cosine similarity for pairs of codes from the same ICD-10 or ICPC-2 chapter, while keeping a low cosine similarity for pairs from distinct chapters (i.e., the difference between same chapter and distinct chapter pairs should be more accentuated) as well as producing clearly defined clusters (i.e., each cluster should only contain data-points of a single chapter) computed through both supervised and unsupervised clustering metrics introduced previously.

One of most important aspects of this work is the interpretability of the results, which is a critical factor in the medical domain, and the reason behind the employment of NMF and variants techniques in our experiments. For all methods experimented, we proceed to generate a topic table with the most important features extracted. For instance for the FR-NMF method, the topics generated by setting $\tau = 5$ are shown in Table 1. Furthermore, we also visualize the relationships between code or clinical terms extracted, for example, in Figure 3 are portrayed the relationships between a subset of ICD-10 codes extracted.

The Figure 3 is created by measuring the cosine similarity between the extracted codes from different chapters, producing interesting relationships. For instance, there is a strong relationship between *accidental falls* (code W199) and *renal insufficiency* complications encoded with (N10, N19, N179, N189). These findings are supported in the medical literature as reported by López-Soto et al. [2015] and Papakonstantinopoulou and Sofianos [2017]. Both teams showed that elderly patients of chronic kidney diseases are more prone to injuries, and with higher probability of fall reincidence. The authors show that other risk factors might be at play including malnutrition, Vitamin-D deficiency and more common occurring in hemodialysis patients. Although of these findings show high correlation, López-Soto et al. [2015] argue that more research is needed in order to precisely calculate the incidence and risk factors of falls in population with chronic insufficiency in order to take preventive measures.

Moreover, the results show that FR-NMF model is able to differentiate between clearly defined causes of death, such as *car accidents* (V892), against other codes which encode health complication or natural causes, only showing similarity with *unspecified intracranial injury* (S069), which might be related to the fact that these are used in combination by a healthcare professional when assigning the codes for the cause of death.

The results for the similarity results for code-pairs and phrase-pairs collections are displayed on Table 2 and Table 3, which displays the mean of cosine similarity for chapters, block and their different both ICD-10 and ICPC-2 classification systems. Let us start by discussing the results achieved for code-pairs collections. From all methods, SNMF achieved highest cosine similarity (i.e., skewed towards the value 1), however this result is common across pairs from same chapter and different chapter. Similarly, DSNMF also inherits the same problems of SNMF, although the values are less skewed towards 1. These results are justified by poor initialization techniques employed by these methods, contrary to NMF-based techniques which use NNDSVDa initialization allowing to create latent representations that better differentiate codes from same chapter or distinct chapter. The best performing methods are not those with higher cosine similarity overall, but those methods that show a higher similarity difference between pairs in the same chapter or distinct chapter (SC - DC). Hence, taking the previous statement into account the best perform-

Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05
insuficincia	cerebral	pneumonia	neoplasia	insuficiencia
respiratria	vascular	j189	pulmo	cardiaca
doena	acidente	aspirao	carcinoma	i509
cardaca	i64	bilateral	c349	insuficiencia cardiaca
insuficincia cardaca	acidente vascular cerebral	pneumonia nosocomial	adenocarcinoma	respiratoria
crnica	acidente vascular	nosocomial	metastizada	renal
diabetes	arterial	comunidade	metastizado	congestiva
cardio	hipertenso	sepsis	mama	cronica
renal	hipertenso arterial	f03	pulmonar	i500
hipertenso	i678	associada	colon	cardio

Table 1: Top ten features in five topics generated by FR-NMF.

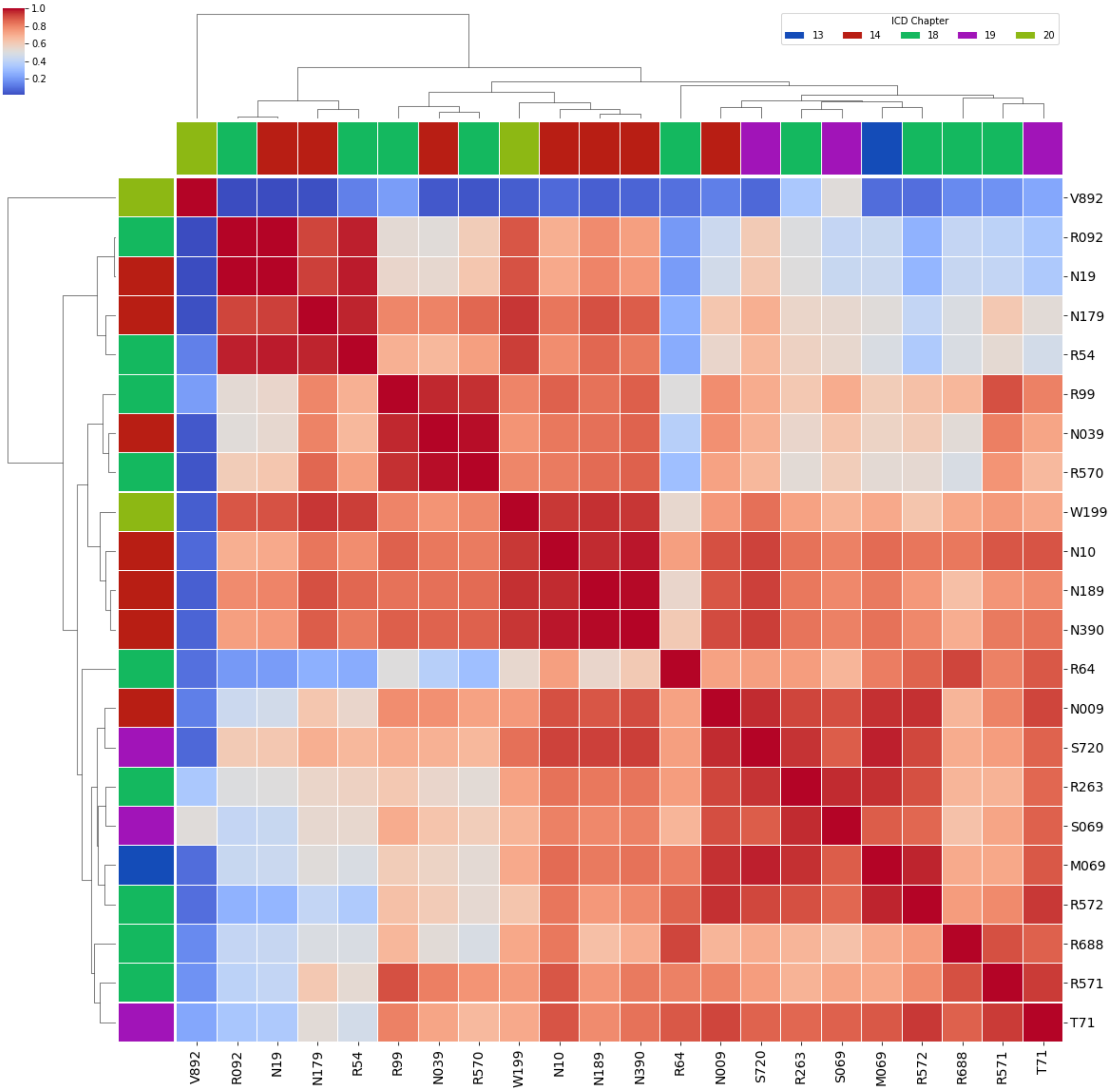

Figure 3: Cosine similarity between codes extracted by NMF taking into consideration all blocks from chapters 13, 14, 18, 19 and 20 of the ICD-10 classification system. Showing the strong correlation between accidental falls (W199) and renal insufficiency complications (N10, N179, N189).

Model	ICD							ICPC		
	SC	DC	SC - DC	SB	DB	SB - DB	SB - SC	SC	DC	SC - DC
FR-NMF	0.342	0.170	**0.172**	0.500	0.187	0.313	0.158	0.180	0.151	0.029
KL-NMF	0.220	0.059	0.161	0.441	0.071	**0.370**	**0.221**	0.110	0.052	**0.058**
SNMF	0.866	0.811	0.055	0.893	0.810	0.083	0.027	0.820	0.818	0.002
DSNMF	0.753	0.754	-0.001	0.758	0.760	-0.002	0.005	0.756	0.765	-0.009
AE	0.779	0.772	0.007	0.776	0.771	0.005	-0.003	0.768	0.759	0.009

Table 2: Average cosine similarity between pairs of ICD-10 and ICPC-2 codes considering $\tau = 25$ latent components with respective difference between pairs.

Model	P ICD							P ICPC		
	SC	DC	SC - DC	SB	DB	SB - DB	SB - SC	SC	DC	SC - DC
FR-NMF	0.243	0.130	**0.113**	0.400	0.148	0.252	0.157	0.377	0.187	0.19
KL-NMF	0.102	0.031	0.071	0.317	0.035	**0.282**	**0.215**	0.366	0.083	**0.283**
SNMF	0.821	0.796	0.025	0.856	0.785	0.071	0.035	0.819	0.801	0.018
DSNMF	0.765	0.756	0.009	0.762	0.755	0.007	-0.003	0.755	0.764	-0.009
AE	0.791	0.785	0.006	0.788	0.784	0.004	-0.003	0.825	0.791	0.034

Table 3: Average cosine similarity between pairs of ICD-10 and ICPC-2 phrases considering $\tau = 25$ latent components with respective difference between pairs.

ing methods are the FR-NMF and the KL-NMF which use the Frobenius norm and the Kullback-Leibler divergence (described in Section 3), because the difference between cosine similarity of ICD SB - SC and ICD SC - DC given by these methods is higher than any other method, indicating that both approaches based on NMF are better at differentiating between codes from same chapters and different chapter, and by analogy better at preserving the hierarchical structure of the ICD-10 classification system. Furthermore, both these methods produced slightly higher cosine similarity for code pairs from the same ICPC-2 chapter (ICPC SC), although the difference (ICPC SC - DC) is not statistically significant because the notches of ICPC SC and ICPC DC boxplot overlap, hence we are left unable to conclude if the ICPC-2 hierarchical structure is preserved by any method based on collections of code pairs.

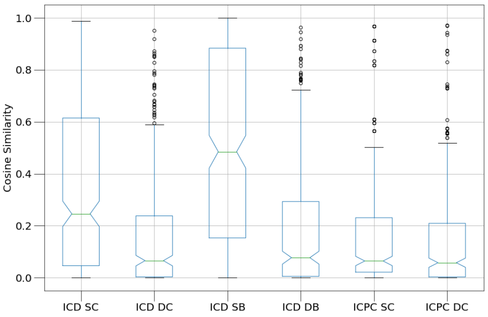

Figure 4: Cosine similarity between pairs of ICD-10 codes produced by FR-NMF.

Now, let us analyze the results achieved for phrase-pairs collection (Table 3), which for the most part are similar to code-pairs collection for the ICD-10 classification system, i.e., codes from the same chapter and block (P ICD SB) produced higher similarity between them than pairs from same ICD-10 chapter (P ICD SC), and these scored higher than pairs from different chapter (P ICD DC) or different block (P ICD DB). For the ICPC-2 classification, there was an improvements on differentiating between same chapter (P ICPC SC) and distinct chapter (P ICPC DC) for FR-NMF and KL-NMF methods. In this case, we can conclude the difference P ICPC SC - DC is statistically significant as the notches of P ICPC SC and P ICPC DC do not overlap as seen in Figure 5. Meanwhile, for SNMF, DSNMF and AE models there was no improvement, producing the nearly the results as when considering collection of code-pairs in Table 2.

To better understand the variability within and between collections created achieved by FR-NMF model, we present the cosine similarity for code-pairs and phrase-pairs collection in Figure 4 and Figure 5, these figures are also used to test if difference between two collections is statistically significant, in case if the notches do not overlap. Hence, for the FR-NMF model only the difference between ICPC SC - DC is not statistically significant.

Figure 6 shows the CV coherence scores across

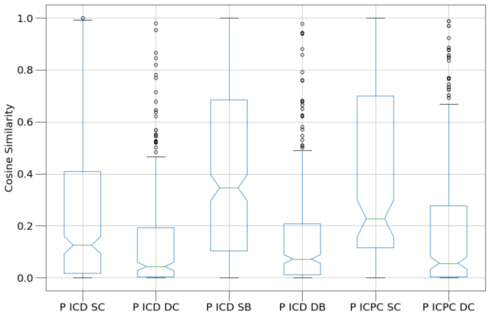

Figure 5: Cosine similarity between pairs of keyphrases produced by FR-NMF.

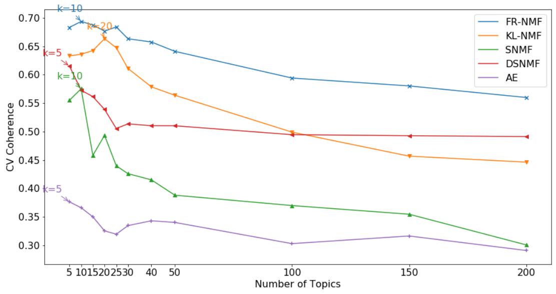

Figure 6: CV Coherence results achieved by each method considering different number of topics τ. A higher coherence score is better.

multiple models implemented. The coherence results obtained align the cosine similarity results, having both methods based on NMF achieve higher coherences scores across all tested topic values τ.

6. Conclusions

The main contributions in this work was the creation of an architecture that uses key-extraction techniques such as AutoPhrase, vector space representations (TF-IDF and Doc2Vec) and dimensionality reduction techniques (NMF, SNMF, DSNMF and AE) that processes a dataset from the medical domain extracting relevant clinical terms present in the data. After extraction and representation we compared each dimensionality technique based on the ability to preserve the hierarchical structure (both intra-chapter and inter-chapter) of two classification system, namely ICD-10 and ICPC-2, showing that only methods based on NMF were capable of preserving the hierarchical structure.

Furthermore, we also measured the performance of all dimensionality reduction techniques from a topic modelling and clustering standpoint. We showed that from a topic modelling perspective the neural networks based dimensionality reduction techniques, such as DSNMF and AE, showed worse results than methods based on NMF. These results are related to the fact that neural network based techniques have been proposed for clustering and reconstruction tasks, in which we also tested, and indeed showed better performance than NMF based methods.

One of the shortcomings of this work is that the results were not validated by medical expert, which would have brought more insight on relationship between codes or clinical terms that share high similarity and evaluate the topics generated by each method employed. Another limitations of this work it that the performance of the models have yet to be tested under a different dataset in order to assess how these perform under different conditions with different content and possible a different classification system for codes used.

For future work we seek to explore unsupervised key-extraction techniques such as YAKE, instead of AutoPhrase, proposed by Campos et al. [2019] and run experiments with other dimensionality reduction methods based on autoencoder extensions which are more closely related to NMF and to topic modelling, e.g., PAE-NMF proposed by Squires et al. [2019] or NVDM proposed by Miao et al. [2016] which was not considered in this work, as these methods can not be compared based on the cosine similarity, because these produce a probability distribution in the latent space and not n-dimensional vectors.

References

Moumita Bhattacharya, Claudine Jurkovitz, and Hagit Shatkay. Identifying patterns of associated-conditions through topic models of Electronic Medical Records. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pages 466–469, 2016.

Christos Boutsidis and Efstratios Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4): 1350–1362, 2008.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alpio Jorge, Clia Nunes, and Adam Jatowt. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2019.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504, 2016.

Chris H. Q. Ding, Tao Li, and Michael I. Jordan. Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:45–55, 2010.

Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9), 2011.

David Gefen, Jake Miller, Johnathon Kyle Armstrong, Frances H Cornelius, Noreen Robertson, Aaron Smith-Mclallen, and Jennifer A Taylor. Identifying Patterns in Medical Records through Latent Semantic Analysis. *Communications of the ACM*, 61(6):72–77, 2018.

Oguzhan Gencoglu. Deep Representation Learning for Clustering of Health Tweets. *arXiv*, abs/1901.00439, 2019.

Geoffrey E. Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *Computing Research Repository*, 2014.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, volume 32, pages 1188–1196, 2014.

Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 535–541, 2000.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. Mining Quality Phrases from Massive Text Corpora. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1729–1744, 2015.

Pablo Jesús López-Soto, Alfredo De Giorgi, Elisa Senno, Ruana Tiseo, Annamaria Ferraresi, Cinzia Canella, María Aurora Rodríguez-Borrego, Roberto Manfredini, and Fabio Fabbian. Renal disease and accidental falls: a review of published evidence. *BioMed Central Nephrology*, 16(1):176, 2015.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, pages 1727–1736, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Computing Research Repository*, abs/1301.3781, 2013.

Konstantina Papakonstantinopoulou and Ioannis Sofianos. Risk of falls in chronic kidney disease. *Journal of Frailty, Sarcopenia and Falls*, 02:33–38, 2017.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 399–408, 2015.

Najibesadat Sadati, Milad Zafar Nezhad, Ratna Babu Chinnam, and Dongxiao Zhu. Representation Learning with Autoencoders for Electronic Health Records: A Comparative Study. abs/1908.09174, 2019.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. Automated Phrase Mining from Massive Text Corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30:1825–1837, 2018.

Steven Squires, Adam Prügel-Bennett, and Mahesan Niranjan. A Variational Autoencoder for Probabilistic Non-Negative Matrix Factorisation. *arXiv*, abs/1906.05912, 2019.

George Trigeorgis, Konstantinos Bousmalis, Stefanos P. Zafeiriou, and Björn W. Schuller. A Deep Semi-NMF Model for Learning Hidden Representations. In *Proceedings of the International Conference on Machine Learning*, volume 32, pages 1692–1700, 2014.

Lance De Vine, Mahnoosh Kholghi, Guido Zuccon, Laurianne Sitbon, and Anthony Nguyen. Analysis of word embeddings and sequence features for clinical information extraction. In *Proceedings of the Annual Workshop of the Australasian Language Technology Association*, pages 21–30, 2015.

Xiao Zhang, Dejing Dou, and Ji Wu. Learning Conceptual-Contexual Embeddings for Medical Text. abs/1908.06203, 2019.

Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang. Measuring patient similarities via a deep architecture with medical concept embedding. In *Proceedings of the IEEE International Conference on Data Mining*, pages 749–758, 2016.