Interpretability on Clinical Analysis from Pattern Disentanglement Insight

Anonymous ACL submission

Abstract

Diagnosis of a clinical condition can help medical professionals save time in clinical decision-making and prevent overlooking risks. Therefore we explore the problem of clinical text interpretability using free-text medical notes recorded in electronic health records (EHR). MIMIC-III is a de-identified EHR database containing observations from over 40,000 patients in critical care units. Since medical notes are free-text, existing machine learning models may have ineffective interpretability; however, interpretability is often desirable for clinical diagnosis. Hence, in this paper, we propose a text mining and pattern discovery solution to discover strong association patterns from patient discharge summaries and the code of international classification of diseases (ICD9 code). The proposed approach offers a straightforward interpretation of the underlying relation of patient characteristics in an unsupervised machine learning setting. The clustering results outperform the baseline clustering algorithm and are comparable to baseline supervised methods.

1 Introduction

If Machine Learning (ML) is to play a significant role in supporting clinical decision making, then it is essential to gain clinician trust (Kim, 2021). Interpretability is frequently defined as the degree to which a human can understand the cause and reason of ML model decisions. The higher the interpretability of a model implies the better the comprehension and explanation of the problem, leading to more accurate and reliable predictions. Most ML algorithms today concentrate on prediction power using general-purpose learning algorithms on large and complex data. However, even though some ML models can also provide various degrees of interpretability, they generally sacrifice interpretability for predictive power (Ghannam and Techtmann, 2021). Therefore, in this study, we focus on interpreting the diagnostic characteristics/patterns from the electronic health records (EHR).

Due to the complex nature of clinical language, clinical texts are often hard to interpret. Topic modeling (Blei et al., 2003) has been applied to the unstructured notes of EHRs to predict clinical outcomes but not with relations to interpretability (Bright et al., 2021; Huang et al., 2015; Wang et al., 2020). Hence, in this study, we use topic modeling to transform free text into interpretable features for pattern discovery and association. In addition, with the recent development in neural networks, variants of pre-trained BERT (Devlin et al., 2018) have widely been applied to clinical domains (e.g. BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019)) with limited interpretability (Feng et al., 2020; Wallace et al., 2019; Van Aken et al., 2021).

Hence, to address the issue of ML interpretability of EHR, we created a novel two-stage algorithm (Figure 1), leveraging interpretable text mining such as topic models (Chen et al., 2019) and pattern discovery techniques (Wong et al., 2021), to discover strong association patterns from patient profiles and discharge summaries to reveal their relationships with the diagnosed disease [1]. The output of the proposed system is an interpretable Knowledge Base, which can link the pattern groups, discovered characteristics of records, and patients' records together to show "what" (disease), "who/where" (tracking patient records back) and "why" (discovered patterns) to interpret clinical notes for better clinical decision making.

To evaluate the performance of the proposed algorithm, we present both a knowledge base with discovered patterns and clustering results. To verify the effectiveness of discovered patterns, we interpret patterns from a clinical perspective to discuss the interpretability of output. As for the clustering results algorithm, although the process of cluster-

[1] ICD9 code, which is the code of international classification of diseases

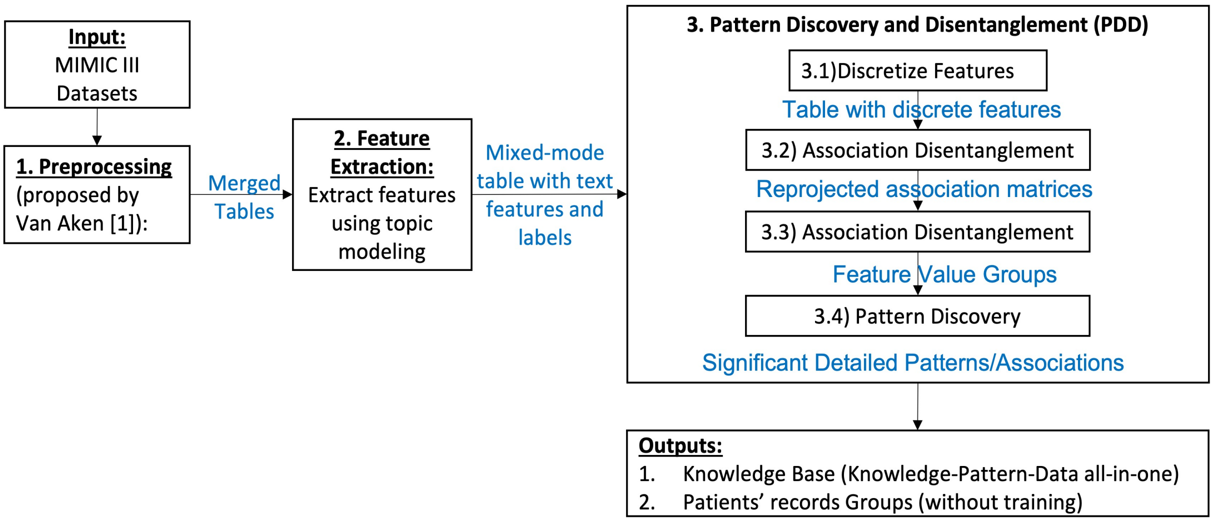

Input:
MIMIC III Datasets
1. Preprocessing (proposed by Van Aken [1]):
Merged Tables
2. Feature Extraction: Extract features using topic modeling
Mixed-mode table with text features and labels
3. Pattern Discovery and Disentanglement (PDD)
3.1)Discretize Features
Table with discrete features
3.2) Association Disentanglement
Reprojected association matrices
3.3) Association Disentanglement
Feature Value Groups
3.4) Pattern Discovery
Significant Detailed Patterns/Associations
Outputs:
1. Knowledge Base (Knowledge-Pattern-Data all-in-one)
2. Patients' records Groups (without training)

Figure 1: The overview of the proposed algorithm

ing records does not require class label information, the results can be evaluated by balanced accuracy and weighted F1- score using the presumed class labels (ICD9 code) as ground truth.

The contributions of the paper are three-fold: 1) Interpretability: a novel algorithm focusing on white-box model interpretation for free-text clinical notes; 2) Unsupervised Disentaglement: the grouping of records based on the discovered associations revealing characteristics of records via unsupervised learning; 3). ALl-In-One Knowledge-Base: generating an all-in-one knowledge base to link knowledge, pattern, and records together for clinical interpretability.

2 Material: MIMIC-III Data Description

MIMIC-III is a de-identified relational clinical database containing observations from over 40,000 patients in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). While MIMIC-III consists of several tabular and time-series datasets, our present study utilizes clinical notes, found in the NOTEEVENTS table, and diagnoses, found in the DIAGNOSES_ICD table.

The former table, NOTEEVENTS, can provide us with the medical notes as text for a detailed description of medical center visits for each patient. The clinical notes contain an internal semi-structured format, which are subdivided into several components, such as: chief complaint, medical history, social history, and discharge information. Each observation refers to a unique hospital stay. The data are related to other tables through unique patient identifiers, hospital stay identifiers, and caregiver identifiers. The latter table, DIAGNOSES_ICD, can provide us with the diagnosis of each patient based on ICD9 codes, which are used as labels to be predicted, and linked with clinical notes.

In summary, our final data contains 11,537 rows/records with the top four classes/diseases represented by ICD9 code. The ICD9 codes are defined as follows: 414 - chronic ischemic heart disease, 038 - septicemia, 410 - acute myocardial infarction, and 424 - diseases of the endocardium. The four classes were slightly imbalanced, with 3502, 3184, 3175, and 1676 observations, respectively, as Figure 2 shows. We chose to include only the top 4 most common codes to highlight the pattern-discerning capability of PDD, as including many codes (especially those with fewer observations) would decrease the interpretability and performance even for supervised learning models.

3 Methodology

In this section, we present the proposed methodology applied to the MIMIC-III dataset. The algorithm proposes tasks in three main steps: preprocessing, feature extraction, and pattern discovery. The overview of the proposed algorithm is shown in Figure 1.

3.1 Preprocessing

We first apply a preprocessing pipeline proposed by Van Aken et al. (2021) to clean and merge the dataset.

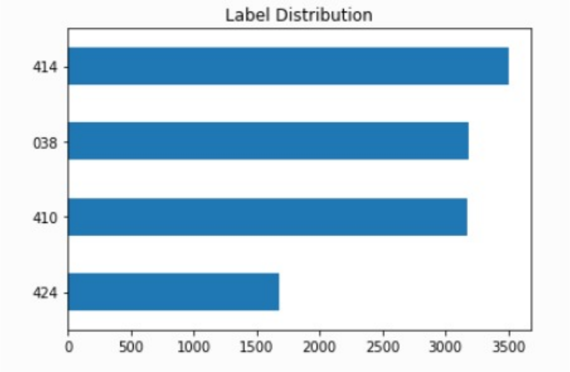


Figure 2: The distribution of labels (Label 414 and Label 424 shows imbalance)

We select the NOTEEVENTS (containing the unstructured text) and DIAGNOSES_ICD tables from the MIMIC-III database. The selected records contain clinical notes of patients, diagnoses, procedures, and ICD9 codes with admission ID columns acting as a link for all the tables. As each admission often had multiple diagnoses, we filter the data by only considering the highest priority diagnosis as the label to be predicted. We then trim the data to the top four most common ICD9 codes.

After retaining these approximate 11,000 text records, we apply regular expressions to remove invalid characters and common stop words as well as words under three characters. We conform every remaining letter to lowercase and apply lemmatization. Finally, we remove a custom list of stop words that are ubiquitous among all text records.

Then, we process the text into a format suitable to be passed as a corpus (embedded lists). A dictionary, or key-value pair, is created from the tokens that were derived from our corpus of cleaned words.

3.2 Feature Extraction

Topic modelling (Hamed Jelodar and Zhao, 2018) is described as a method for finding a group of words (i.e topic) from a collection of documents that best represents the information in the collection. Hence, we extract features from the clean dataset using topic modelling. The value of the features represented by the probabilities of topics occurring in the records. Labels are then merged with the features for unsupervised exploration; in this case, the label is the ICD9 code - the diagnostic code indicating categories of disease. We use LDA (Latent Dirichlet Allocation) for the topic model because it identifies topics best describing distinct subsets of documents within a corpus (Hamed Jelodar and Zhao, 2018).

To determine the ideal number of topics, we choose the optimal number of topics by computing coherence of the topic cluster instance (Röder et al., 2015). We find that the coherence score peaks when the number of topics is 5, 20, and 30 - and therefore we create topic models with those respective parameters. The output of our coherence scores is shown as Figure 3.

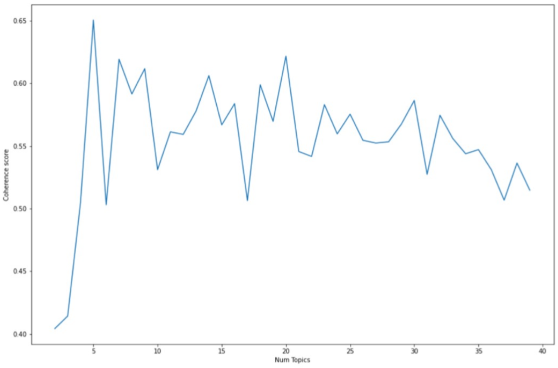


Figure 3: Optimal number of topics by coherence of the topic cluster

3.3 Pattern Discovery and Disentanglement

After preprocessing and extracting features from the text, the dataset has been transformed into a structured table of patients' records in rows and features in columns, which is represented as a $M \times N$ matrix, where M represents the number of patients' records and N represents the number of extracted features [2].

3.3.1 Discretize Numerical Feature Values

The output matrix in the last step contains probabilities of topics or extracted words, which are all numerical values. Due to infinite degrees of freedom of numerical features, it is hard to correlate features with the target variable and interpret the associations. Hence, we discretize features into event-based/discrete features. To detect event-based patterns, we convert the values of numerical features into categorical features by using the Equal Frequency discretization which distributes the values into equal size bins, so that numerical feature values are converted into discrete values referred to as "feature value" (meaning the discrete value for that feature). To be consistent with the study of

[2] In pattern discovery, we use the term attribute instead of feature.

PDD (Wong et al., 2021), we use the term Attribute Value (AV) instead.

3.3.2 Association Disentanglement

In order to measure the association between a pair of AVs (i.e. certain values of one attribute co-occurs with the value of another attribute), we use the statistical measure of adjusted standardized residual, abbreviated by SR, to represent the statistical weights of the AV pair, which is denoted as SR($AV_1 \leftrightarrow AV_2$) (shorten as SR($AV_{12}$)) and calculated by Eqn. (1) below.

$$SR(AV_{12}) = \frac{Occ(AV_{12}) - Exp(AV_{12})}{\sqrt{Exp(AV_{12})}} \times (1 - \frac{Occ(AV_1)}{T}\frac{Occ(AV_2)}{T}) \tag{1}$$

where $Occ(AV_1)$ and $Occ(AV_2)$ are the number of occurrences of AV; $Occ(AV_{12})$ is the total number of co-occurrence for two AVs in a AV pair; and $Exp(AV_{12})$ is the expected frequency and T is the total number of records.

An association matrix, treated as a vector space, is then generated to represent the strength of associations between each pair of AVs. Each row of the matrix, corresponding to a distinct AV, represents an AV-vector with SRs between that AV associated with all other AVs corresponding to the column vectors as its coordinates. We call the matrix the SR Vector Space (SRV). SRV is an *N* dimensional vector space consisting of *N* distinct AV-vectors.

We then use PCA to decompose SRV (Wong et al., 2021) (Wong et al., 2018) into principal components to reveal AV associations orthogonal to others AV associations, i.e. PC=PC_1, PC_2,... PC_k which are ranked according to the weights of the associations (eigenvalues). We then reproject the projections of AV-vectors on the principal components onto the SRV again, to obtain a set of reprojected-SRVs (abbreviated by RSRV). We refer to the PC together with its RSRV as a disentangled space.

The above process is called *Pattern Disentanglement* which allows us to take the reprojected components/vectors from PCA and use the reprojected values as new measurements/criteria to represent the strength of associations between AVs in different orthogonal disentangled spaces.

3.3.3 Obtain Attribute Value Groups with Disentangled Associations

In an RSRV, after screening in the statistical residual values (referred to as RSR) greater than 1.96, only the significant pairs of AV associations remain. Statistically, under the null hypothesis that the two AVs are independent, the adjusted residuals will have a standard normal distribution. So, an adjusted residual that is more than 1.96 (2.0 is used by convention) indicates the association is significantly greater than what would be expected (with a significance level of 0.05 or 95% confidence level) if the hypothesis were true. We can also set a threshold as 1.44 with 85% confidence, or 1.28 with 80% confidence level.

As an unsupervised learning approach, on each RSRV, we generate AV groups such that each group contains a set of AVs. We build the set of AVs up iteratively by adding AVs that are associated with AVs in the set. That is to say an AV (e.g., AV_i) that is significantly associated with another AV (e.g. AV_j) in the group will join the group, otherwise, a new AV group is generated for AV_i. Theoretically, in one projected principal component, usually two AV groups on the opposite sides are generated as two opposite groups. When such opposite groups do not exist, we may obtain AV groups only on one side of the PC. The output of this step is one or two AV groups, and each group contains a set of AVs.

Furthermore, to obtain detailed separated groups, several AV subgroups can be generated for each AV group using a similarity measure such that the similarity between two AV subclusters is specified as the percentage of the overlapping records covered by each AV subcluster. We denote each AV subgroup by a three-digit code [#PC, #Group, #SubGroup]. The AV groups or subgroups can reveal the characteristics of the records at specific groups with disentangled patterns to provide statistical evidence for further clustering or prediction. Furthermore, patient record groups are obtained according to their specific characteristics (disentangled patterns) discovered in the AV groups or subgroups.

3.3.4 Pattern Discovery on Attribute Value SubGroups

Traditional pattern clustering algorithm (Zhou et al., 2016), without PCA, can group patterns based on their "similarity", which is limited and time-consuming. In this case, after disentanglement and generating AV groups/subgroups, only a few AVs remain to be candidate patterns, which can reduce time consumption when high-order patterns are growing. The high-order pattern describes a statistically significant association among more than two AVs.

So far, each AV subgroup contains a set of AVs considered as candidate patterns. We then test the candidates from order > 2 (i.e. consisting of more than 2 AVs) to high order sets to determine their pattern status. Hence, we obtain a compact set of patterns which are statistically significant and interpretable. Hence PDD reduces the computational complexity drastically and produces very small and succinct pattern sets for interpretation and tracking. The disease related record groups of patients can then be explicitly revealed.

3.4 Output

The output of PDD is organized into an all-in-one representational framework known as PDD Knowledge Base. It consists of three parts: a Knowledge Section showing the hierarchical clusters such that each cluster unveil distinct characteristics of a related group of records; a Pattern Section listing the discovered patterns showing detailed associations between AVs; and the Data Section listing the record ID's, the knowledge source and pattern(s) associated with each patient by linking the patient to the Knowledge and Pattern Sections

4 Experimental Result

We present our results in Table 1 and knowledge base in Figure 4 and Figure 5.

4.1 Topic Modeling Result

From a clinical perspective, the generated topic models correspond reasonably well with each ICD9 diagnosis. In the 20-topic model, septicemia - a widespread infection of the body, was predicted by topics containing relevant words such as "infection", “bacteria", and "culture". Conversely, topics that contained cardiovascular-related terms such as "ventricular" or "aorta" predicted the heart-related diagnoses. Additionally, the algorithm was able to discern the heart-related diagnoses from one another: dividing acute myocardial infarction (410) from the more chronic and congenital diseases (414, 424). The algorithm may have discerned that words representing severe prognoses or procedures, such as "angioplasty", "emergency", and "death" were more correlated with acute myocardial infarction. Taken together, topic modeling and PDD provides an interpretable methodology to predict ICD9 diagnosis with reasonable accuracy when given unstructured clinical text as input.

4.2 Comparison of Unsupervised and Supervised Learning

Although the process of clustering individuals does not require class label information, the entity clustering performance can be evaluated from the clustering results by two statistical measures using the presumed class labels as ground truth. In this study, since the numbers of records belonging to different classes are imbalanced, the correct prediction of the majority classes will overwhelm that of the minority classes. In this case, we followed the same evaluation method in (Van Aken et al., 2021), *balanced accuracy* (Balanced Acc. in Table 1) and *weighted F1-scores* (Weighted F1 in Table 1), to evaluate performance of both supervised and unsupervised results. Balanced accuracy is defined as the average of recall obtained in each class (Brodersen et al., 2010) and the weighted F1-score is calculated by averaging the support-weighted mean per class F1-scores (i.e. weights on class distribution) (Chakravarthi et al., 2020). Both above results are referred to the *sklearn.metrics* package in Python 3.0 (Pedregosa et al., 2011).

We compared the clustering results of PDD with the classical clustering algorithm, K-mean, as the baseline, and also two supervised learning algorithms: Random Forest (Breiman, 2001) and CNN (Kalchbrenner et al., 2014). The data were split into 70% training and 30% for testing.

As for K-means, we use the *sklearn.clusters* package in Python 3.0 (Pedregosa et al., 2011) with all default parameter settings and assign the number of clusters as four. For Random Forest, we apply the default parameter settings from the package of *sklearn.ensemble.RandomForestClassifier* in Python 3.0 (Pedregosa et al., 2011).

For CNN (LeCun et al., 1995), we trained a CNN model with the input layer as a reshaped cleaned dataset with probabilities of topics or extracted words and ICD9 labels. The architecture is as follows: a 1D CNN layer, followed by batch normalization, then a dropout layer for regularization (Li et al., 2019), and finally a 1D max-pooling layer. After the CNN and pooling, the learned features are flattened to one long vector and passed through a fully connected layer before the output

Unsupervised Learning								
Features	$TFIDF_{40}$		TM_5		TM_{20}		TM_{30}	
Algorithms	K-mean	PDD	K-mean	PDD	K-mean	PDD	K-mean	PDD
Acc.	0.49	0.50	0.59	**0.78**	0.72	0.72	0.58	0.70
Balanced Acc.	0.48	0.45	0.62	**0.78**	0.74	0.74	0.51	0.73
Precision	0.48	0.75	0.58	**0.84**	0.73	0.73	0.50	0.73
Recall	0.49	0.45	0.62	**0.78**	0.74	0.74	0.51	0.73
Weighted F1	0.42	0.41	0.57	**0.78**	0.72	0.72	0.56	0.71
Avg. F1	0.44	0.38	0.57	**0.78**	0.71	0.71	0.50	0.70
Supervised Learning								
Features	$TFIDF_{40}$		TM_5		TM_{20}		TM_{30}	
Algorithms	RF	CNN	RF	CNN	RF	CNN	RF	CNN
Acc.	0.82	**0.84**	0.66	0.67	0.74	0.72	0.74	0.73
Balanced Acc.	0.81	**0.85**	0.62	0.62	0.72	0.70	0.71	0.70
Precision	0.82	**0.84**	0.64	0.67	0.74	0.72	0.74	0.73
Recall	0.81	**0.84**	0.62	0.67	0.71	0.72	0.71	0.73
Weighted F1	0.82	**0.84**	0.65	0.66	0.74	0.72	0.73	0.72
Avg. F1	0.82	**0.84**	0.63	0.67	0.72	0.72	0.72	0.73
AUC.	0.95	**0.96**	0.87	0.88	0.91	0.90	0.91	0.91

Table 1: Experimental Result Comparison.

layer for prediction. We used Adam optimizer with a learning rate of 0.001 trained on 25 epochs with a batch size of 32.

As the baseline comparison for features, we also applied all supervised and unsupervised learning algorithms on the dataset with words extracted using TFIDF (Jones, 1972). In a corpus, frequent words in one document tend to be frequent in all other documents. TFIDF (term-frequency-inverse document frequency) is an algorithm that scores words that are distinctively frequent in a particular document but not necessarily within the general corpus. TFIDF can be computed as:

$$\text{tf-idf(t,d)} = \text{tf(t,d)} \times \text{idf(t)}$$

where tf refers to the term frequency (proportion of a particular term t over all terms); and

$$\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$$

where n is the total number of documents in the set and df is the number of documents containing the term t.

To discover associations among features and class labels and to make the interpretation meaningful, we did not keep all words in TFIDF, but selected the top 40 words with a feature selection algorithm by Random Forest.

The comparison results are shown in Table 1. It is interesting to observe that PDD outperformed other models but underperformed when applied on the TFIDF results, which consist of the results of K-means. Both supervised learning algorithms, Random Forest and CNN perform better on the TFIDF dataset. The reason should be that the top 40 words (feature) are selected based on classification results.

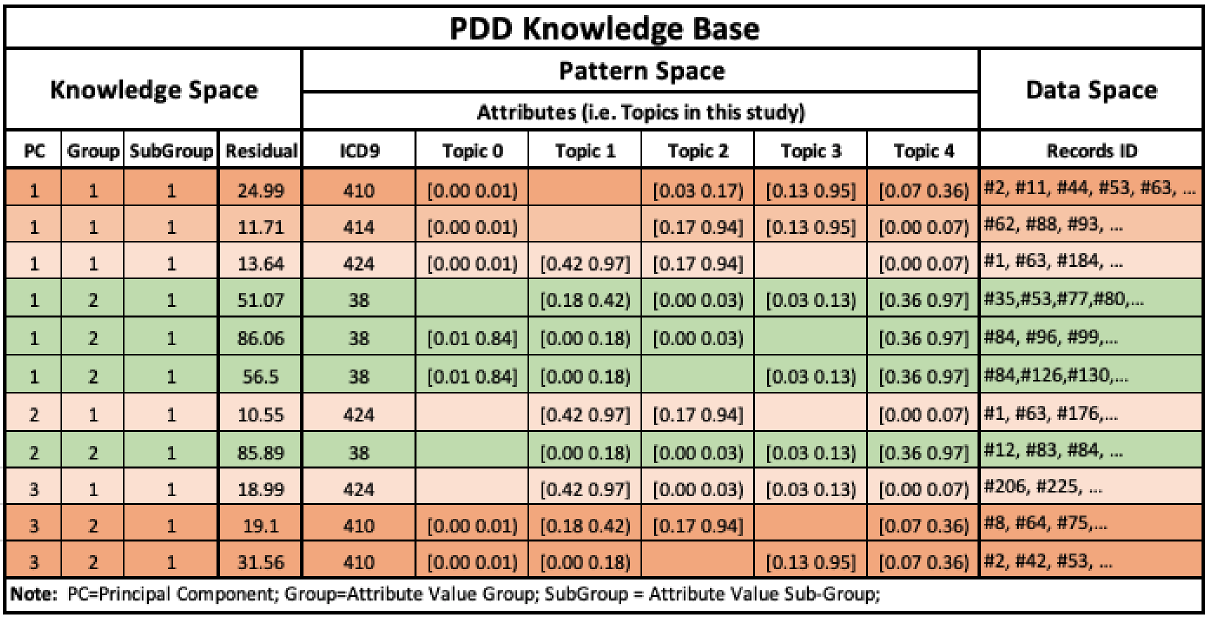
When topic modeling results are used as a dataset, PDD outperforms K-means and even the two other supervised learning algorithms, with balanced acc.=0.78 and weighted F1-score=0.78, when only 5 topics are used. As for Random Forest, it performs better when applied to the topic modelling results with 20 topics than another the two experiments running on 5 topics and 30 topics. While as for CNN, the results of experiments on 30 topics are slightly better than the results on 20 topics.

One important notion we would like to bring forth is that, even if the accuracy score reflects the algorithm performance to some extent, class labels may not always be reliable in supervised classification algorithms. On the contrary, clustering merely recognizes patterns in the data and holds no such risk.

4.3 Interpretability

From the perspective of interpretability, when the topic modeling dataset with top 5 and top 20 topics were compared, the clustering performance of PDD is superior to all the other methods. As an example, we present the PDD Knowledge Base on 5 topics and 20 topics as shown in Figure 4.

The first three columns show the knowledge space, which are clustering results of PDD and statistical measurement of each pattern. The clusters are identified by a three-digital code [#PC, #Group, #Subgroup] (PC: Principal Component,



PDD Knowledge Base

Knowledge Space				Pattern Space						Data Space
				Attributes (i.e. Topics in this study)						
PC	Group	SubGroup	Residual	ICD9	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Records ID
1	1	1	24.99	410	[0.00 0.01)		[0.03 0.17)	[0.13 0.95]	[0.07 0.36)	#2, #11, #44, #53, #63, ...
1	1	1	11.71	414	[0.00 0.01)		[0.17 0.94]	[0.13 0.95]	[0.00 0.07)	#62, #88, #93, ...
1	1	1	13.64	424	[0.00 0.01)	[0.42 0.97]	[0.17 0.94]		[0.00 0.07)	#1, #63, #184, ...
1	2	1	51.07	38		[0.18 0.42)	[0.00 0.03)	[0.03 0.13)	[0.36 0.97]	#35,#53,#77,#80,...
1	2	1	86.06	38	[0.01 0.84]	[0.00 0.18)	[0.00 0.03)		[0.36 0.97]	#84, #96, #99,...
1	2	1	56.5	38	[0.01 0.84]	[0.00 0.18)		[0.03 0.13)	[0.36 0.97]	#84,#126,#130,...
2	1	1	10.55	424		[0.42 0.97]	[0.17 0.94]		[0.00 0.07)	#1, #63, #176,...
2	2	1	85.89	38		[0.00 0.18)	[0.00 0.03)	[0.03 0.13)	[0.36 0.97]	#12, #83, #84, ...
3	1	1	18.99	424		[0.42 0.97]	[0.00 0.03)	[0.03 0.13)	[0.00 0.07)	#206, #225, ...
3	2	1	19.1	410	[0.00 0.01)	[0.18 0.42)	[0.17 0.94]		[0.07 0.36)	#8, #64, #75,...
3	2	1	31.56	410	[0.00 0.01)	[0.00 0.18)		[0.13 0.95]	[0.07 0.36)	#2, #42, #53, ...

Note: PC=Principal Component; Group=Attribute Value Group; SubGroup = Attribute Value Sub-Group;

Figure 4: The PDD Knowledge Base for 5 topics are used as input.

Group: pattern groups in the same principal component, Subgroup: pattern Sub-group in the same pattern group). We observe that, in the first principal component, two opposite groups are discovered: one where ICD9=4XX, and the other where ICD9 = 038. All ICD9=4XX are diseases related to heart disease, while ICD9=038 is related to Septicemia, so these are two opposite groups. Then in the second principal components, ICD9=424 (diseases of the endocardium) was separated, still showing opposite patterns with ICD9=38. Finally, in the third principal component, ICD9=424 was separated from ICD9=410 (acute myocardial infarction).

Then, the pattern space shows the discovered significant associations between ICD9 code and the extracted topics. To be more specific, the unveiled knowledge can be summarized as below.

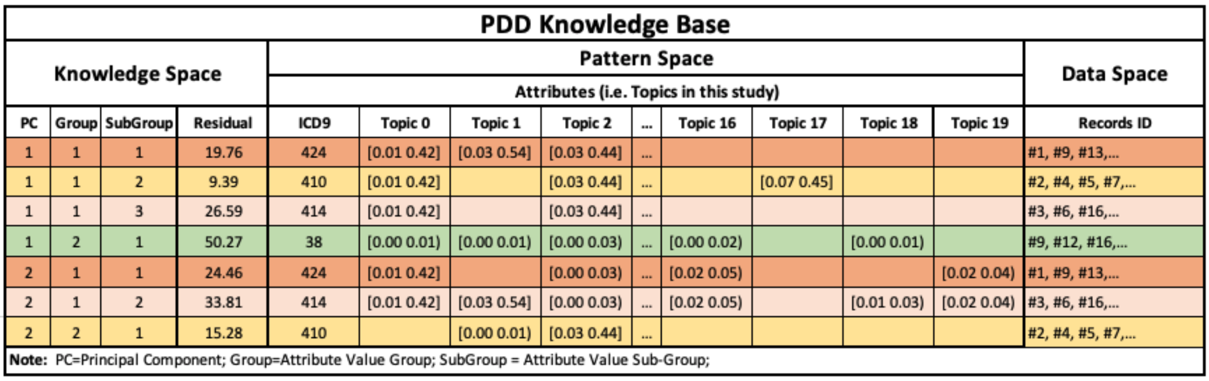
- ICD9=424,410,414 (heart diseases) show similar patterns with Topic 0 (Medication) showing low probabilities.
- ICD9=424 (endocardium disease) and 414 (chronic ischemic heart disease) show more closed patterns compared to 410 (acute myocardial infarction), topic 4 (Intensive Care/Infection) showing low probability. And the unique characteristic of ICD9=424 (endocardium disease) is that Topic 1 (Cardiovascular 1) showing high probability.
- ICD9=38(septicemia) shows opposite characteristics compared to ICD9=4XX, with Topic 0 (Medication) showing high probability, Topic 2 (Cardiovascular 2) showing low probability, and Topic 4 (Intensive Care/Infection) showing high probability.

The data space shows the records IDs that are covered by the patterns. For example, the first association pattern listed in the first row of the knowledge base can be covered by the records with ID = 2,11,44,53,63 and so on. And all above records belong to the group labeled as ICD9=410, which is same with the discovered pattern.

In addition, Figure 5 shows the partial knowledge base on 20 topics dataset. As same with the above results, in the first principal component, two opposite groups are discovered: one where ICD9=4XX (heart diseases), and the other where ICD9 = 038 (septicemia). But the difference is that three subgroups (i.e. 424, 414, 410) are detected related to three different ICD9 codes in the first group in the first principal component.

Similar to the above results using 5 topics, the discovered significant patterns can be summarized for 20 topics as below. Since the most of topics are not clear, we highlighted the meaning for partial topics.

- ICD9=424 (diseases of the endocardium) and 414 (chronic ischemic heart disease) shows similar patterns, for example: i) **high** probabilities appear in the topics 1,2(Car-



PDD Knowledge Base													
Knowledge Space				Pattern Space									Data Space
				Attributes (i.e. Topics in this study)									
PC	Group	SubGroup	Residual	ICD9	Topic 0	Topic 1	Topic 2	...	Topic 16	Topic 17	Topic 18	Topic 19	Records ID
1	1	1	19.76	424	[0.01 0.42]	[0.03 0.54]	[0.03 0.44]	...					#1, #9, #13,...
1	1	2	9.39	410	[0.01 0.42]		[0.03 0.44]	...		[0.07 0.45]			#2, #4, #5, #7,...
1	1	3	26.59	414	[0.01 0.42]		[0.03 0.44]	...					#3, #6, #16,...
1	2	1	50.27	38	[0.00 0.01)	[0.00 0.01)	[0.00 0.03)	...	[0.00 0.02)		[0.00 0.01)		#9, #12, #16,...
2	1	1	24.46	424	[0.01 0.42]		[0.00 0.03)	...	[0.02 0.05)			[0.02 0.04)	#1, #9, #13,...
2	1	2	33.81	414	[0.01 0.42]	[0.03 0.54]	[0.00 0.03)	...	[0.02 0.05)		[0.01 0.03)	[0.02 0.04)	#3, #6, #16,...
2	2	1	15.28	410		[0.00 0.01)	[0.03 0.44]	...					#2, #4, #5, #7,...

Note: PC=Principal Component; Group=Attribute Value Group; SubGroup = Attribute Value Sub-Group;

Figure 5: The PDD Knowledge Base when Top 20 topics are used as input.

diovascular/Surgery),5,16; ii) and topics with **low** probabilities are topics 6, 7 (Status/Consciousness), 8 (Lung disease), 9.

- ICD9=038 (septicemia) shows opposite patterns, for example: i) topics with **high** probabilities are topics 3, 4 (Intensive care/Infection), 7 (Status/Consciousness), 8 (Lung disease); ii)and **low** probabilities appear in the topics 0(Heart anatomy) 1, 2 (Cardiovascular/Surgery), 5, 12 (Cardiovascular), 16, 18.

Compared to more simple methods of interpretability such as extracting feature importance from random forest, one can use PDD to interpret the feature associations with particular groups of interest. For example, feature importance from the random forest model for 20 topics ranks topics 5, 6, and 4 as highest importance while 14, 15, 16 as lowest importance based on impurity. While a healthcare practitioner might understand which topics are useful for differentiating between diagnoses with feature importance rankings, they cannot gain further insight into the associations between features and groups: associations that would be necessary to support a diagnosis. Therefore, PDD offers a more functional level of interpretability compared to feature importance.

5 Conclusion

In this work, we propose a novel two-step algorithm, using interpretable NLP features with unsupervised pattern discovery to solve clinical text analysis. Experiments show results from both clustering accuracy and interpretability. As for the clustering results, PDD performs better than K-means, especially when applied to the dataset extracted by topic modeling. Clustering results of PDD based on the discovered patterns may reflect the functional sources of the original dataset instead of class labels.

MIMIC III is a clinical dataset de-identified in accordance with HIPAA's Safe Harbor guidelines, where 18 patient identifiers such as name, address, dates, and telephone number are removed (Johnson et al., 2016). However, when given additional information or with linkage to other public datasets, re-identifying patients from even de-identified data is possible (Khaled El Emam and Malin, 2011) (Latanya Sweeney and Brody, 2017). While we expect that the patient data is generally well-protected, powerful predictive and pattern-discovery algorithms may amplify an undue risk of re-identification. Thus, one must regulate the use of algorithms to ensure that the patient data is used for its intended purpose.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Roselie A Bright, Summer K Rankin, Katherine Dowdy, Sergey V Blok, Susan J Bright, and Lee Anne M Palmer. 2021. Finding potential adverse events in the unstructured text of electronic health care records: Development of the shakespeare method. *JMIRx Med*, 2(3):e27017.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010*

20th international conference on pattern recognition, pages 3121–3124. IEEE.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

Jinying Chen, John Lalor, Weisong Liu, Emily Druhl, Edgard Granillo, Varsha G Vimalananda, and Hong Yu. 2019. Detecting hypoglycemia incidents reported in patients' secure messages: using cost-sensitive learning and oversampling to reduce data imbalance. *Journal of medical Internet research*, 21(3):e11990.

Phil Culliton, Michael Levinson, Alice Ehresman, Joshua Wherry, Jay S Steingrub, and Stephen I Gallant. 2017. Predicting severe sepsis using text from the electronic health record. *arXiv preprint arXiv:1711.11536*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable clinical decision support from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1478–1489.

Tushaar Gangavarapu, Aditya Jayasimha, Gokul S Krishnan, and Sowmya Kamath. 2020. Predicting icd-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems*, 190:105321.

Ryan B Ghannam and Stephen M Techtmann. 2021. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84.

Chi Yuan Xia Feng Xiahui Jiang Yanchao Li Hamed Jelodar, Yongli Wang and Liang Zhao. 2018. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78.

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15(1):55–86.

Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. 2017. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, 12(4):e0174708.

Zhengxing Huang, Wei Dong, and Huilong Duan. 2015. topic model for clinical risk stratification from electronic health records. *Journal of Biomedical Informatics*, 58:28–36.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:16.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Efsun Sarioglu Kayi, Kabir Yadav, and Hyeong-Ah Choi. 2013. Topic modeling based classification of clinical reports. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 67–73.

Luk Arbuckle Khaled El Emam, Elizabeth Jonker and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS One*, 6.

Been Kim. 2021. Interpretability.

Matt D Price Kimberly Raiford Wildes John F Hurdle Kimberly J O'Malley, Karon F Cook and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health Services Research*, 40.

Laura Perovich Katherine E. Boronow Phil Brown Latanya Sweeney, 1 Ji Su Yoo and Julia Green Brody. 2017. Re-identification risks in hipaa safe harbor data: A study of data from one environmental health study. *Technology Science*.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. 2019. Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2690.

Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, Wim Vanden Berghe, Bart Goethals, and Kris Laukens. 2015. A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231.

Miha Pavlinek and Vili Podgorelec. 2017. Text classification method based on self-training and lda topic models. *Expert Systems with Applications*, 80:83–93.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. *arXiv preprint arXiv:2102.04110*.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.

Yanshan Wang, Yiqing Zhao, Terry M Therneau, Elizabeth J Atkinson, Ahmad P Tafti, Nan Zhang, Shreyasee Amin, Andrew H Limper, Sundeep Khosla, and Hongfang Liu. 2020. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *Journal of biomedical informatics*, 102:103364.

Andrew KC Wong and Gary CL Li. 2008. Simultaneous pattern and data clustering for pattern cluster analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(7):911–923.

Andrew KC Wong, Ho Yin Sze-To, and Gary L Johanning. 2018. Pattern to knowledge: Deep knowledge-directed machine learning for residue-residue interaction prediction. *Scientific reports*, 8(1):1–14.

Andrew KC Wong, Pei-Yuan Zhou, and Zahid A Butt. 2021. Pattern discovery and disentanglement on relational datasets. *Scientific reports*, 11(1):1–11.

Pei-Yuan Zhou, Gary CL Li, and Andrew KC Wong. 2016. An effective pattern pruning and summarization method retaining high quality patterns with high area coverage in relational datasets. *IEEE access*, 4:7847–7858.

A Materials and Methods

An EHR is a digital collection of medical information about a person, which includes information about a patient's health history, such as diagnoses, medicines, tests, allergies, immunizations, and treatment plans. The MIMIC-III (Medical Information Mart of Intensive Care) is an openly available extensive database comprising de-identified information relating to patients admitted to critical care units at a large tertiary care hospital (Johnson et al., 2016). Data primarily stores both structured (e.g. MIMIC-III medications, laboratory results are stored in the table with columns as features and rows as records) and unstructured data (e.g. MIMIC-III clinical notes, discharge summaries are stored in the format of free text). The discharge summary of patients is free text, thus making interpreting it a challenge.

The first step is transforming free text into a structured dataset formatting as a table with columns as features and rows as records. The second step is discovering patterns and grouping patients' records based on patterns in an unsupervised manner.

B Limitations

This study has the following limitations. First, to prove the concept of the PDD algorithm, only records with the four most common ICD9 codes are selected. Second, PDD, used as an interpretable clustering algorithm in this study, accepts limited selected features. When too many features are included, acquired data leads to high time complexity, and overwhelming pattern number and redundancy, making interpretability very difficult. For future work, we will enlarge the dataset and the number of features to investigate their impact on the performance of the algorithm. Finally, as the predicted label is ICD9 code, we presume it to be ground truth for diagnosis. However, ICD9 is used for billing purposes and therefore may not accurately reflect a patient's true condition (Kimberly J O'Malley and Ashton, 2005).

C Related Work

C.1 Clinical Data Analysis with Interpretability

Due to the complex nature of clinical language, clinical texts were hard to interpret. Most of the previous works on clinical data analysis were based on structured data, which lack complementary information such as lab reports or patient history. Clinical expert judgments may thus require information that are available only in unstructured data (e.g. clinical texts) (Culliton et al., 2017).

Latent Dirichlet Allocation (LDA)(Blei et al., 2003) topic modeling (Blei et al., 2003) has been applied to the unstructured notes of EHRs to predict clinical outcomes (Bright et al., 2021; Huang et al., 2015; Wang et al., 2020). In addition, Ghassemi et al. (2014) showed latent topic features were more predictive than structured features, and a combination of the two performs best.

Topic features cluster terms into a small set of semantically related groups, which is proved useful in text classification and categorizing clinical reports (Chen et al., 2019; Pavlinek and Podgorelec, 2017; Kayi et al., 2013). For example, Horng et al. (2017) combined structured and unstructured data for sepsis prediction using text modeling involving topic models. Further, Gangavarapu et al. (2020) proposed a vector space and topic modeling-based approach applied to structure the raw clinical data by exploiting the data in the nursing notes. Hence, in this study, we use topic modeling to transform free text into a table with features and records.

In addition, with the recent development in neural networks, variants of pre-trained BERT (Devlin et al., 2018) have widely been applied to clinical domains (e.g. BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019)). In addition, Feng et al. (2020) used pre-trained BERT-based models as static feature extractors and showed that variants of BERT performed better with Sepsis than Mortality prediction tasks however Wallace et al. (2019) showed that BERT fails to interpret life-threatening important numerical values such as body temperature in the clinical text. Further, Van Aken et al. (2021) showed that medical-specific negations can be misinterpreted by the pre-trained language models such as BERT (e.g. "abstinence from alcohol" becomes "alcohol dependence syndrome"). Finally, BERT does not provide interpretable features. Unlike topic models or TF-IDF, a BERT vector does not contain any explicit semantic information that can be easily interpretable by a person.

C.2 Pattern Discovery

To tackle the interpretability of clinical data analysis, many machine learning algorithms were proposed. For example, the Decision Tree can generate a rule set between features and class labels for interpretable prediction, but the rules need to be trained relying on labeled classes. In addition, Frequent Pattern Mining (Naulaerts et al., 2015) (Han et al., 2007) can discover knowledge in the form of association rules from relational data (Han et al., 2007) (Van Aken et al., 2021) but a manually threshold need to be set for calculated likelihood, support or confidence (Van Aken et al., 2021). And the discovered patterns may be overwhelmed (Wong and Li, 2008) with overlapping/redundant patterns, which requires some post analysis approaches, such pattern pruning and pattern summarizing (Wong and Li, 2008).

Hence, in this study, to interpret association between text and class labels, we utilized topic modelling (Chen et al., 2019) with a novel pattern discovery and disentanglement (PDD) algorithm (Wong et al., 2021), which has not been applied in text mining. By using the PDD, it can discover simple patterns with statistical support to reveal the association between extracted features with class labels without further pattern pruning or pattern summarization. The output patterns are well organized, more clear and easier to be comprehended in a knowledge base.