Project 2 - Representation Learning

Ali Ander Berkay Berabi Imre Kertesz

1 Task 1 - Binary Classification

1.1 Data Exploration by Visualization

In order to explore and analyze the data, it was visualized regarding different aspects. As stated in the slides, the data contains many missing values and therefore, a plot highlighting the percentage of missing values was created. Additionally, the readmittance rate with respect to the gender and race was plotted. Both plots can be seen in Figure 1. One observes that weights, payer-code and medical speciality features have a considerable amount of missing values, whereas race and all three diagnosis almost do not miss any value. Additionally, readmittance rate between male and female patients of all races except asian patients are in general very close to each other; asian female patients readmit themselves approximately 20% more than the male asian patients.

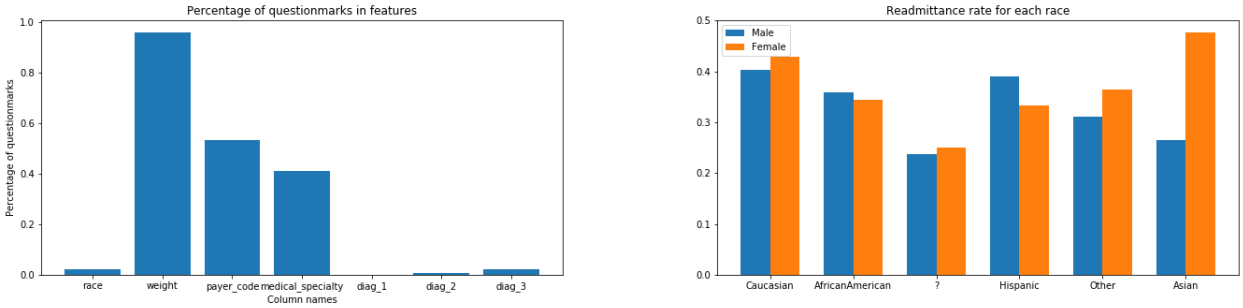


Figure 1: Left: Percentage of Missing Values, Right: Readmittance Rates

Plotting the readmittance rate against the number of procedures (number of lab tests performed during the encounter) shows that patients with a higher number of procedures are less likely to be readmitted. This observation applies to both genders. Looking at the right plot in figure 2 one can see that patients who only stayed one day in the hospital have a low readmittance rate. In the first week of stay a gain in the readmittance rate can be detected. As the time spent in the hospital increases it seems that the readmittance rate for male patients decreases again.

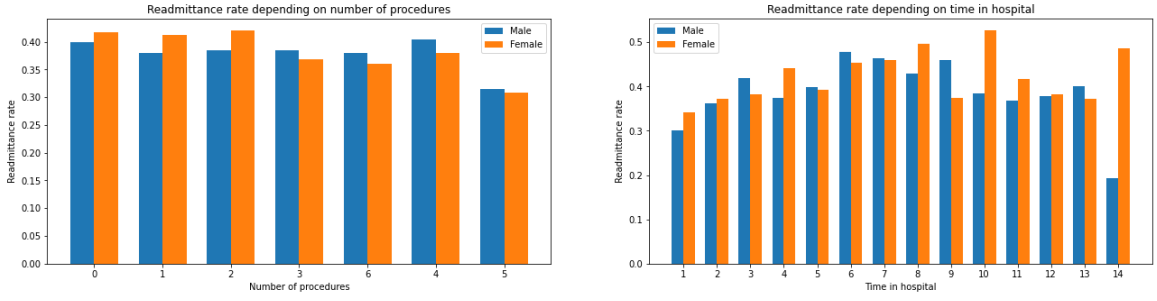


Figure 2: Left: Readmittance rate depending on number of procedures, Right: Readmittance rate depending on time in hospital

Using two more visualization techniques, namely tSNE and word cloud visualization,provided invaluable information about the data. First, the given data points were embedded to 2D by applying tSNE dimensionality reduction algorithm, that preserves the neighborhood relationships and the embeddings were colored regarding their readmission class. The resulting plot can be seen in Figure 3.

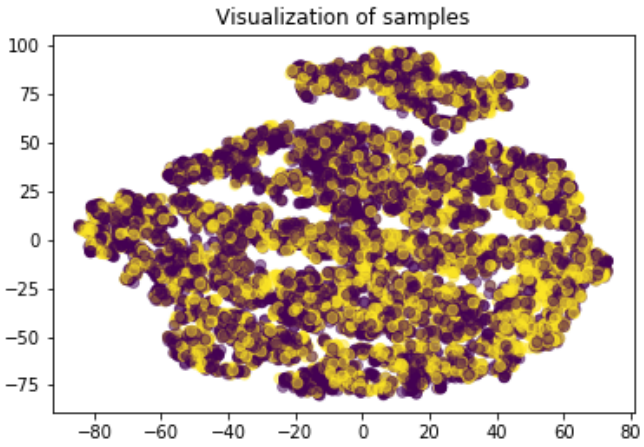
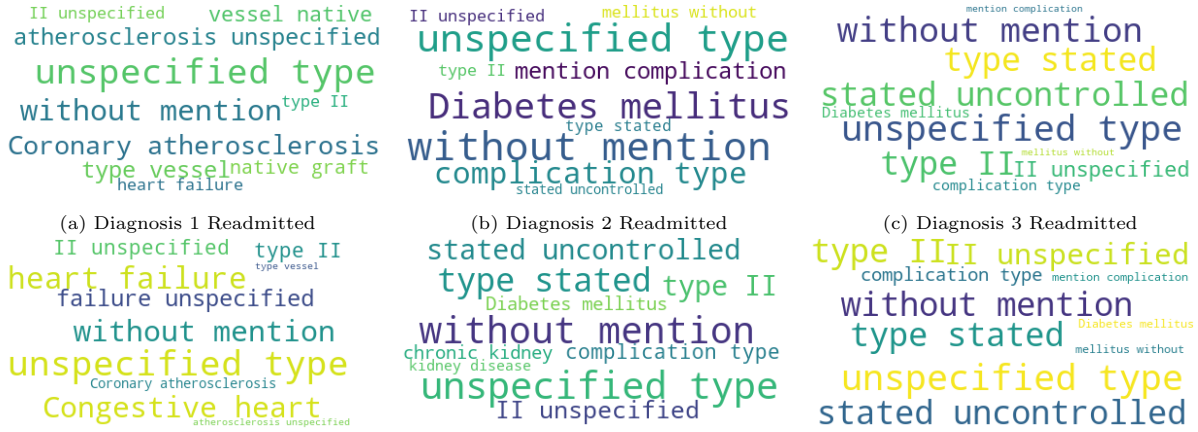


Figure 3: tSNE visualization colored with target class labels

As depicted in the above figure, the data points in 2D embedding are hardly separable. They are roughly scattered through the space and therefore, there seems to be no obvious clusters. This suggests that the provided features are not discriminative at all and classifying points with the given dataset would be a hard task to achieve.



(a) Diagnosis 1 Readmitted

(b) Diagnosis 2 Readmitted

(c) Diagnosis 3 Readmitted

(d) Diagnosis 1 Not Readmitted

(e) Diagnosis 2 Not Readmitted

(f) Diagnosis 3 Not Readmitted

Figure 4: Word Clouds for each class and description

.

Lastly, we generated a word cloud for each class from the textual description of the diagnosis. It is notable that all the word clouds contain the words "without mention" and "unspecified type" together with other common words. This indicates that the diagnosis are not specifying patients' situation well enough. The acquisition of the dataset is therefore questionable and thus, any inference made from this dataset would be not totally reliable.

1.2 Preprocessing Models for Textual Descriptions

In order to solve task 1, various models were tested. The primary goal was to find out whether processing textual data with natural language processing techniques may help improving the performance and to analyze their effect on the model. To this extent, different approaches to extract features from the given textual data were used. Below we explain the preprocessing steps applied for every model. We briefly mention methods we used for feature extraction and furthermore, the final model trained for the classification task.

Before describing the preprocessing models, how the missing values are handled, is explained. We extracted patients' weights and payercodes from the dataset. They are not informative indicators regarding the readmission probability and a huge portion, 97%, of patients' weights are missing. For all the other missing values, we decided to substitute them with appropriate labels representing "missing". (e.g. -1 for numerical features and empty text for textual features). Moreover, following the suggested

categorization of ICD codes in [1], the provided ICD codes were grouped into 9 categories depending on the type of the disease, which allowed us to have a more compact representation of ICD codes.

Baseline Model: Our baseline model does not profit from any natural language processing methods. Textual features are simply removed from the data. This model is used for highlighting the effects of NLP models in this binary classification task.

TF-Model: In this model, we vectorize the text with Term Frequency method and then choose k best features for the sake of dimensionality and compactness.

TF-IDF-Model: In this model, we vectorize the descriptions with Term Frequency Inverse Document Frequency method.

Bigram-Model In this model, we vectorize the text regarding the bigrams. We try two distinct vectorization, one with TF approach and another with TFIDFs.

Trigram-Model: In this model, we vectorize the text regarding the trigrams. We try two distinct vectorization, once with TF approach and once with TFIDFs.

Word2Vec: We apply word2vec model to the textual descriptions with an embedding dimension of 50 and for each patient we calculate the mean vector from the words occuring in their diagnosis.

In all methods, the stopwords were removed and lemmatization was applied while extracting features. For each model except Word2Vec, we apply feature selection with SelectKBest method from scikit-learn framework for the sake of compactness and dimensionality reduction. The processed textual features are combined with the baseline features and fed into a XGBoost Model. For hyper parameter tuning we apply grid search with 4 folds giving 2000 samples for validation. Lastly, we predict the readmittance of patients from the test dataset and evaluate the models by their F1-Score and AUROC.

1.3 Results

According to the experiments' evaluations, we conclude that none of the aforementioned NLP models led to a noticeable improvement of performance metrics compared to the baseline model. In what follows, we will explain the general and model-specific reasons for the lack of performance improvement.

We think, the reason why the NLP models did not outperform the baseline model is that the corresponding ICD codes of provided textual descriptions are already extant in the original data. These codes represent the categorical information of the descriptions, for example, the ICD code 250 corresponds to Diabetes mellitus. Every other extension of the ICD code 250 such as 250.0, 250.1 ... etc. is a type of Diabetes and specified as "Diabetes mellitus without mention of complication" and "Diabetes with ketoacidosis " respectively. Therefore, making use of various NLP models to extract features from a structured information set is not effective as the descriptions do not incorporate any new valuable information about patient's illness. Moreover, the descriptions do not provide any observed patient specific information such as symptoms. This could be a valuable source for discriminating the patients from each other since it is a well known fact that the effects and symptoms of an illness on each patient varies immensely. We believe that one could achieve more valuable insights from textual data, if the descriptions were more detailed and patient specific rather than a general illness description.

Lastly, we were only provided 10000 samples, while the original data set contains 100.000 samples. With more data, it would also be possible to increase the performance.

TF Model It is actually not surprising that this method did not provide an increase in performance, since it is the most basic NLP method. TF ignores word orders and all terms are considered equally important. Using tf based features, the AUROC score was increased by 0.0008 while the f1-score was decreased by 0.0010. Therefore, we can say that TF-model did not help at all.

TF-IDF Model TF-IDF solves the problems of TF by scaling down the weights with respect to document frequency. Therefore, TF-IDF is able to extract more important words from the model and in fact performs also better than TF model. TFIDF increased the AUROC score as well as the F1-score of the baseline model by 0.082 and 0.083 respectively.

Bigram and Tigram Models Both TF and TF-IDF suffers from the fact that they can not capture the order of the words since they use a bag of word approach. Therefore, N-grams performs in general better than them. However, this was not the case for our experiments. The results from bigram and trigram models were pretty close. We think that this was the case due to the fact that the words in our corpus are extremely specific and unique, appearing in a very limited context, for instance, pericarditis, hypertension, mellitus, atherosclerosis, prostate can be mentioned as examples. Most of the words are medical terms and thus, they are very specific. Using N-grams to consider their relationship with other surrounding words does not contribute to model's performance due to the fact that these words appear only in a specific domain of disease and have similar neighboring words, e.g., the word mellitus only appear with the word diabetes. The n-gram models also provided some improvements with respect to the baseline model, but actually none of them performed better than TFIDF-model. Among ngram models, trigram-tf is the best model, having 0.6486 as AUROC and 0.6052 as f1-score.

Word2Vec Word2vec method was not able to prove any valuable insight to the baseline model, as the performance metrics even decreased.

All results are summarized in the following Table 1.

	Baseline	*TF*	*TFIDF*	*Bigram-TF*	*Trigram-TF*	*Bigram-TFIDF*	*Trigram-TFIDF*	*Word2Vec*
AUCROC	0.6441	0.6449	**0.6523**	0.6368	0.6486	0.6477	0.6456	0.6355
F1-Score	0.5993	0.5983	**0.6076**	0.5922	0.6052	0.6035	0.6002	0.5729

Table 1: Area under Receiver Operator Characteristic Curve, F1-scores and accuracy of the models, the maximal values for every row is marked bold

2 Task 2 - Information Retriaval from CORD 19 dataset

2.1 Approach

Our first approach is to create a general model to solve any task that takes queries as input, retrieves most relevant papers being highly correlated with the query and outputs sentences containing the most informative keywords in those papers. Below you find an explanation about the pipeline.

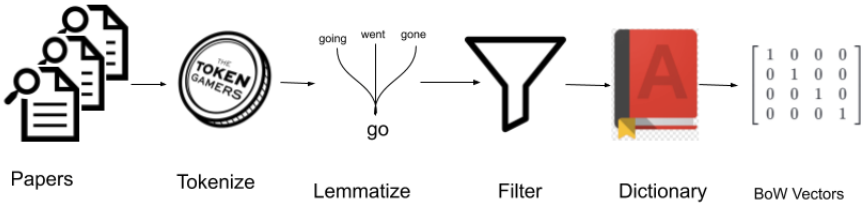


Figure 5: Pipeline of preparing the BoW vectors

Our first approach only focuses on the papers that are published in years either 2019 or 2020, since the first COVID-19 case appeared in 2019. This of course reduces the dataset size to be processed from 47000 to 5329 papers. As seen in the Figure 5, we tokenize each of these papers and lemmatize them using the wordnet lemmatizer from gensim framework. We filter out the punctuations and stopwords predefined in gensim. Following that, a dictionary is built and every paper is transformed into a BoW vector according to the dictionary.

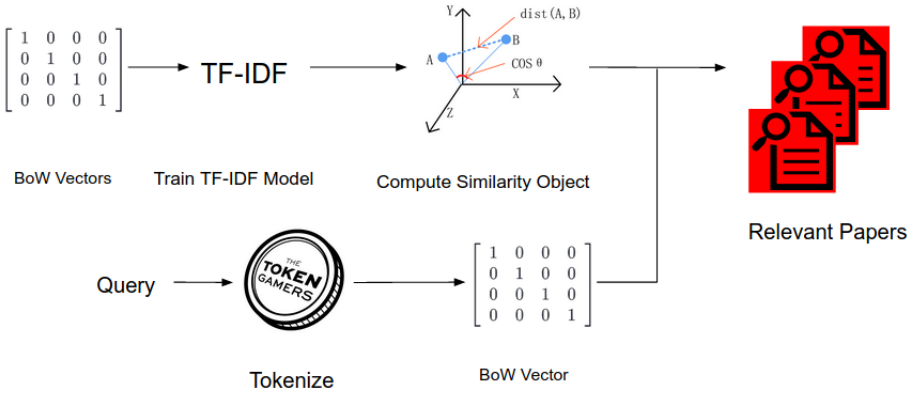


Figure 6: Querying from the BoW Vectors

After computing the BoW vectors of this corpus, we compute the similarity matrix with the tf-idf transformed vectors, After receiving a query, the query will be transformed into a document with tf-idf transformed BoW vectors and the most similar K papers to the query are selected from the dataset according to the cosine similarity measure.

Assuming that the steps above success in their task of retrieving relevant papers, the selected papers should contain information about the query. Therefore, the most important information of the paper should contain the most relevant information about the query. Following this logic, we find the most important words in the document by selecting a certain amount of words in the paper that got the highest scores in tf-idf transform. Afterwards, we search for the sentences in which these words occur and return them as answer to the query.

Two results for the query "risk factor corona virus 2019" are given below:

" Indeed, RA patients show increased infectious risk because of impairment of immune system and immunosuppressive related-therapy"

" The patients are diverse, and the medical staff is at risk of infection"

References

[1] Beata Strack, Jonathan Deshazo, Chris Gennings, Juan Luis Olmo Ortiz, Sebastian Ventura, Krzysztof Cios, and John Clore. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed research international*, 2014:781670, 04 2014.