Automated classification of cardiology diagnoses based on textual medical reports

J. A. O. Pedrosa, D. M. Oliveira, Wagner Meira Jr. and Antonio Luiz P. Ribeiro

Universidade Federal de Minas Gerais, Brazil
{joao.pedrosa, derickmath, meira}@dcc.ufmg.br and tom@hc.ufmg.br

Abstract. Automatic classification of diagnoses has been a long term challenge for Computer Science and related disciplines. Textual clinical reports can be used as a great source of data for such diagnoses. However, building classification models from them is not a trivial task. The problem tackled in this work is the identification of the medical diagnoses that are indicated in these reports. In the past, several methods have been proposed for addressing this problem, but a method developed for reports in the cardiology area that are written in Portuguese is still needed. In this paper we describe a method that is able to handle the peculiarities of clinical reports, including the medical terminology, and that is implemented to estimate correctly the diagnosis based on raw clinical reports and a list of the possible diagnoses. Experimental results show that our method has a high degree of accuracy, even for infrequent classes and complex databases.

Categories and Subject Descriptors: Applied computing [**Life and Medical sciences**]: ; Computing methodologies [**Natural Language Processing**]:

Keywords: cardiology, information extraction, machine learning, natural language processing

1. INTRODUCTION

Descriptive medical reports have been widely used for the development of health-related studies and technologies, which, for instance, extract information organized as a category taxonomy. A key information that is usually present in such medical reports is the set of symptoms and possible diagnoses, but such information may be still limited w.r.t. diagnoses categories and may not support the expression of nuances [Stein HD 2000]. As a consequence, free text analysis is commonly chosen as a strategy when no category precisely describes clinical findings, or when there is a need to give supporting evidence for diagnosis or suspicion [Ford et al. 2013]. In summary, retrieving the diagnoses from a medical report is not a trivial task.

The problem addressed in this work is the categorization of these reports, according to the diagnoses described by them. Given that the number of reports available is usually very large, it means that reviewing them manually is time consuming and cannot be performed in reasonable time for most applications [Paixao et al. 2018], justifying the need for an automated solution.

This problem may be solved with the use of *Natural Language Processing* (*NLP*) methods and models, a technology that has been used for many years [Hripcsak et al. 1995] and its effectiveness has been already proven. Most implemented medical NLP systems reach a recall in the range of 80 - 85% and a precision in the range of 95 - 99% [Mamlin et al. 2003]. Even though this is not a perfect performance, it may be good enough to be used in real-world applications, since humans fall within

The authors would like to thank FAPEMIG, CNPq and CAPES for their financial support. This work was also partially funded by projects MASWeb, EUBra-BIGSEA, INCT-Cyber, ATMOSPHERE and by the Google Research Awards for Latin America program.

the same range.

While there are several models that make use of *NLP* to retrieve and use reports' information [Friedman et al. 1995; Dang PA 2008], the problem relies on the fact that many of them target just the English language or are not developed for the cardiology area, which makes them far from ideal to be applied to our problem scenario.

In this paper, we propose a self-supervised classification model that is specifically developed for clinical reports written in Portuguese. An earlier version of this model, as well as a subset of the results, was presented at KDMILe 2020 and this paper is an extension of the one that was presented at the conference [Pedrosa et al. 2020]. Our method estimates a label from each textual report, indicating what are the diagnoses that are contained in the report. Our strategy does not demand any manual categorization of textual reports for the method to work. In summary, we have created a robust and effective method, and applied it to two real cardiology datasets provided by Hospital das Clinicas de Minas Gerais, one of which comprises more than 2,000,000 reports.

2. PROBLEM DESCRIPTION

The problem we are handling in this paper is a multiclass and multilabel classification task of free-text for medical reports. Usually, medical reports are written as free-text by physicians and retrieving the diagnoses in the text is very important, but, on the other hand, is not an easy task since there is a large amount of data and there are several ways in which the same diagnosis may have been described in the text. The common approach for such classification is a supervised learning method, in which a subset of the existing data is labeled by a specialist and used for building a classifier, but manually labeling the text in this task is too hard. Therefore, we need to build a method that can estimate the diagnosis expected by the doctors without manually labeling.

Since the same diagnosis can be expressed in several different ways, a first resource for implementing our proposal is a *Diagnoses Dictionary*. The Diagnoses Dictionary is elaborated by a specialist and lists how a given diagnosis can be written. For instance, the diagnosis *"Left Atrial Enlargement"* may be also identified as *"Left atrial hypertrophy"* or even *"left atrial abnormality"*. Each one of these terms should be associated with the same entry in the Diagnoses Dictionary. By creating the Diagnoses Dictionary we expect to have enough information about the target diagnosis without manually labeling the data, but, since the medical report is free-text, it is not possible that the Diagnoses Dictionary contains every form that the diagnosis may be found in medical reports. For example, in the diagnosis *"Left Atrial Enlargement"*, we also find *"L.A. Enlargement"*, *"left atrial abnorm"*, *"L.A. abnorm"*, among other variations and potential misspelling.

In this work we evaluate our approach using two different cardiology-related datasets, which are described in the next subsections. In both cases, the only manually labeled exams are those that compose the test dataset.

2.1 Electrocardiogram Records Database

The first dataset we used, named as *Electrocardiogram (ECG) Records Database*, consists of 2,322,513 clinical reports from ECG records of 1,676,384 different patients from 811 counties in the state of Minas Gerais/Brazil. This dataset was acquired through the Telehealth Network of Minas Gerais (TNMG) [Alkmim et al. 2012] and is composed of 68 ECG diagnoses together with a Diagnoses Dictionary, containing common terms used for each diagnosis, which was developed along with a cardiologist.

For the sake of a better understanding of how the dataset looks like, we show next a record sample that illustrates a record containing multiple diagnoses:

> QRS axis shift to the left. QRS: LBBB morphology and increased ST and T wave duration: secondary changes to QTcLBBB: impaired Conclusion: 1- Tachycardia suggestive of marked sinus. 2-Left bundle branch block with left anterosuperior divisional block morphology. 3-Secondary alterations of ventricular repolarization. 4-Isolated Ventricular Extrasystoles with conduction aberration. [...]

This record indicates the diagnoses **Left Bundle Branch Block, Ventricular Extrasystoles, Left Anterior Fascicular Block, Left Axis Deviation, Secondary Changes in Ventricular Repolarization and Sinus Tachycardia**.

To exemplify how diverse each diagnosis may be, the "Ventricular extrasystoles" can be referred in the text as "Ventricular extrasystoles", "EEVV", "Ventricular Ectopy", but it is not limited to these keywords, since this is a free text dataset. It is worth mentioning that, in this scenario, keyword searching is not enough for retrieving the diagnosis, as we discuss later when evaluating the "regular expression" baseline.

2.2 Pacemaker Patients Database

The Second Dataset, named as *Pacemaker Patients Database*, contains records from pacemaker patients. This dataset was acquired through the Telehealth Network of Minas Gerais (TNMG) and is composed of 70,312 records from 2,899 patients from Hospital das Clinicas de Minas Gerais. This dataset is composed of 10 electrocardiography diagnoses and a Diagnoses Dictionary, containing common terms used for each cardiovascular diagnosis, which was developed along with a cardiologist.

3. RELATED WORKS

Several works emerged recently aiming to automatically classify medical textual data. Despite the large volume of data associated with healthcare applications, a significant portion of their data is free text, and does not contain clear patterns that maybe exploited by an automated method. The usage of *NLP* for such problems was proposed by several works [Spyns 1996; Friedman 1997; Souza et al. 2014; Hassanpour and Langlotz 2016] in recent years, but it is still a big challenge.

There are previous works that have used classic text extraction and model building to classify health related data in the fields of radiology and cardiology [Xu and Sharma 2019; Jagannatha and Yu 2016]. These works employed Recurrent Neural Networks, based on Long Short Term Memory cells, and Word Embedding, achieved good results, and have showed that these are effective tools in the realization of this task. On the other hand, they require manual labeling of a large amount of data to support some supervised learning algorithm to classify the reports, and, as it is not always possible to build a training database by manually labeling records, this approach is not applicable to most applications.

Along with that, there are works that use a semantic approach. Some works explore information retrieval models and techniques and use language characteristics to improve the result [Friedman et al. 1995; Spyns 1996], while others explore statistical parsing models [Collins 1997; Klein and Manning 2003]. Statistical parsing models are generative models of lexicalized context-free grammars and it has been shown that these models may be expanded to handle sub-categorization and wh-movement[1]. Results show that this approach can achieve good performance, reaching more than 85% for both accuracy and recall. However, despite the good results, these models are designed for the English

[1]Wh-movement is the formation of syntactic dependencies that make some words, in the English language, change position in a phrase, depending whether the phrase is interrogative or affirmative. Interrogative forms are known within English linguistics as wh-words, such as **wh**at, **wh**en, **wh**ere, **wh**o and **wh**y.

language and, since the information retrieval and statistical parsing are based on characteristics of the language, it is very hard to apply them directly to other languages.

There are three other methods that can be used to address the same problem and that we will use as baselines for comparative assessment to our proposal: Regular Expressions, Latent Dirichlet Allocation (LDA), and Transformer Models [Vaswani et al. 2017].

The classical version of the LDA provides topic modeling in a non-supervised context, since LDA works by connecting each document to each word through a thread, based on their location in the document, and then use this information to learn which documents discuss the same topic. Some works [Yadav 2017; Allahyari et al. 2017] have used a modified version of the Latent Dirichlet Allocation to achieve the task of retrieving information from medical reports. They have used semantic terms in order to achieve good results, and there are works that extended it towards a self-supervised version.

Transformer models were originally developed to address the problem of sequence transduction, or neural machine translation, and are based solely on attention mechanisms [Bahdanau et al. 2014], a procedure that searches for parts of a given relevant sentence to predict a target word in an encoder-decoder model. Thus, they can be applied to practically any task that transforms an input sequence to an output sequence, including speech recognition, text-to-speech transformation and other labeling tasks, and have shown high effectiveness in all of these tasks in previous works [Hu and Singh 2021; Karita et al. 2019; Gulati et al. 2020].

The main difference from our approach to existing ones is that, through the Diagnoses Dictionary, we are able to use the medical knowledge already known from literature, instead of using a completely unsupervised method, while, at the same time, not having to perform a laborious manual classification of the textual reports.

4. METHODOLOGY

In this section we describe our method and discuss each of its components. As mentioned, it is a self-supervised method, since it learns classification patterns from the medical reports using the Diagnoses Dictionary. In particular, each diagnosis from the Dictionary is a class and we will refer to them as classes from now on. We also adopt the premise that our proposed method must use only the database with the textual clinical reports and the Diagnoses Dictionary(Sec. 2), which contains the most common terms for each class, meaning that no manual labeling of the data is necessary. Our method is illustrated in Figure 1. In order to accomplish the classification task, our method stacks three different steps, described in the next subsections.

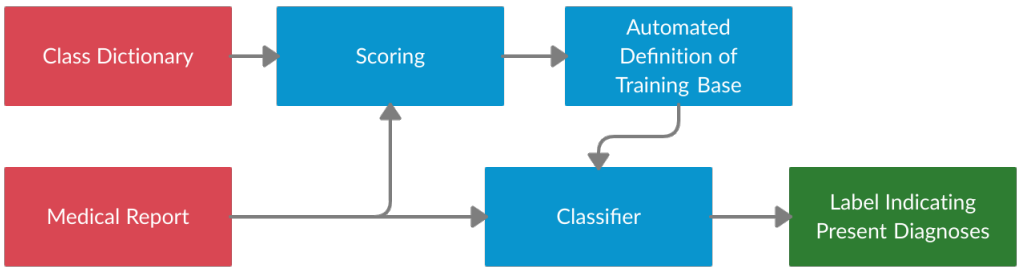

Fig. 1. Illustration of each step in the proposed method. Here, the red rectangles represent the information that is used as input to the method, the blue rectangles represent the method steps and the green rectangles represent the method's outcome.

4.1 Scoring

Scoring aims to quantify the strength of the relation between each textual report and each class. We use the Diagnoses Dictionary that contains the common terms for each diagnosis and the medical report as input for scoring. The entries of the Diagnoses Dictionary may contain either *Normal sentences* or Acronyms, we estimate the *score* for each entry for each class. The *score* is a real number between 0 and 1. The higher the value of the *score* is, the higher is the probability that the text belongs to the desired class. For determining the score for acronyms we employ a binary function, that is, if the acronym is a substring of the text report, then the score is 1, otherwise, it is 0.

For *normal sentences*, the score is defined as a similarity between the sentence and each substring of the medical report. A common similarity measure for text is the Levenshtein distance [Pradhan et al. 2015] and it was used in our method to measure the score for *normal sentences* as follows:

We denote $lev(A, B)$ as the Levenshtein string distance between two strings A and B. Also, lets define a score function f between two strings as:

$$f(A, B) = \frac{max(length(A), length(B)) - lev(A, B)}{max(length(A), length(B))}$$

Denoting A as a term of our dictionary and S as the set that contains all substrings of the clinical report that has the same length of A, the score between this clinical report and the class associated with A is:

$$\max\{f(A, B) : B \in S\}.$$

In summary, the score between a report and a class is the maximum score for all terms of the class present in the Diagnoses Dictionary .

4.2 Automated Definition of Training Base

Using the scores calculated, we generate a training dataset for a classifier. This training dataset is automatically generated by our method and is a subset of the complete database. Another information that we know from medical literature is the prevalence for each class, i.e., the proportion of exams in which each class usually appears. Thus, the dataset consists of the records associated with the highest scores, which should also satisfy a threshold x, chosen so that the number of records that score higher than x is as close as possible to the number of records indicated by the class prevalence. For example, suppose that our dataset has n records and there is a class c with a defined prevalence of 0.01. We would sort the records in descending order of score and the threshold for class c would be the score of the record ranked $\lfloor 0.01 \cdot n \rfloor$. This threshold is depicted in Figure 2.

We can now use the training dataset to build the machine learning model, as described next.

4.3 Classifier

We implemented our classifier using a recurrent neural network to learn the patterns. Our hypothesis is that the training dataset contains latent features so that a supervised learning method can learn from them and then classify the entire dataset.

First, we need to preprocess the medical reports, transforming them into vectors that can be used as input for a classifier. For that, we used a text vectorization model based on frequency in the text inspired by *term frequency – inverse document frequency (tf-idf)* [Ramos et al. 2003].

With the texts now transformed into vectors, we can use these vectors and the training dataset generated in 4.2 as the input for training the classifier. After the learning process, the classifier should

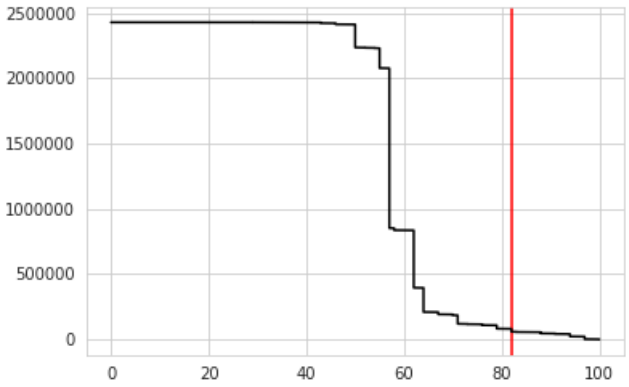

Fig. 2. Visual explanation of how the threshold is chosen. Axis Y represents number of registers that have a score greater or equal to the score defined in Axis X. The red line shows where the threshold must be placed for the number of registers to be as close as possible to the amount defined by the class prevalence.

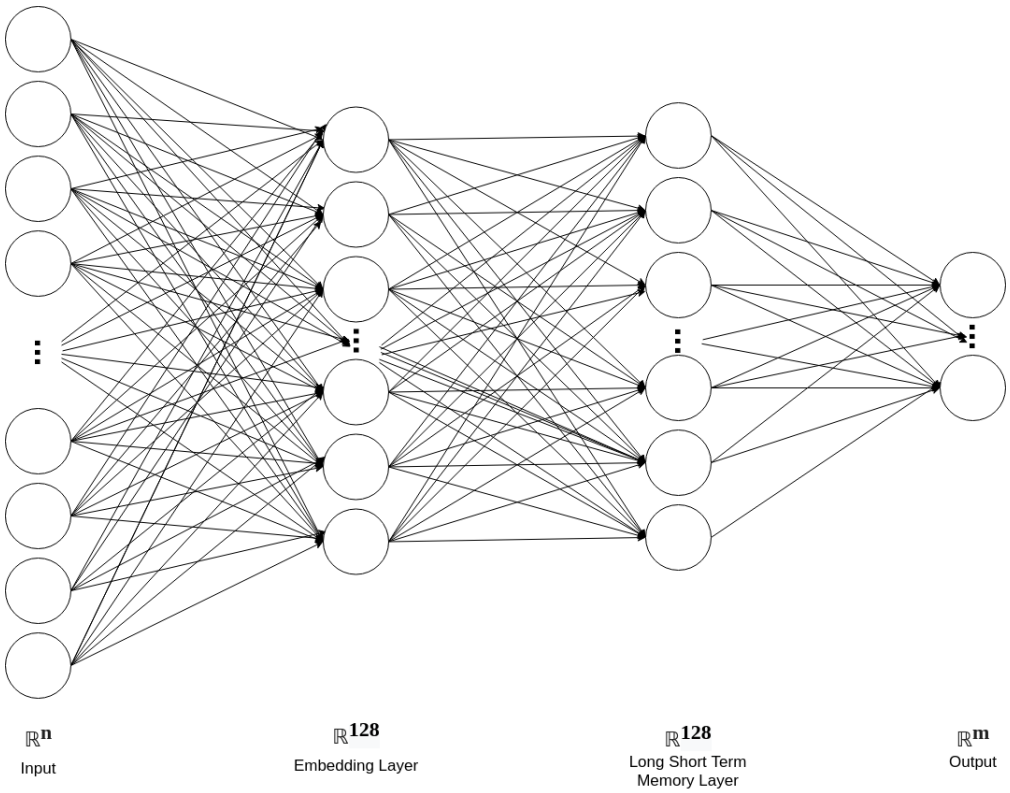

Fig. 3. Structure of the Neural Network.

be able to take a textual report as input and output a label that represents the classes contained in the report.

As mentioned, our classifier is a Recurrent Neural Network model, consisting of two main layers.

The first layer is a word embedding layer. Word embedding is a technique that consists of denoting

semantically similar words [Mikolov et al. 2013]. Relying on the hypothesis that linguistic items with similar distributions have similar meanings [Harris 1954], the technique defines similarity based on the context where the words appear. As a consequence, we can set the word embedding as a parameter in our model, and let it be updated during training.

The second layer is a Long Short Term Memory layer. Long Short Term Memory layers define a special kind of recurrent neural network, capable of learning long-term dependencies. They were introduced in 1997 [Hochreiter and Schmidhuber 1997] and work tremendously well on a large variety of problems.

After the network has been built, the information gathered in the process described in the last two steps is used to train it.

During the training step, the loss function used was the *Binary Cross Entropy Loss*. The binary cross entropy loss is defined as:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))))$$

where N is the number of records, y is the label (1 if belonging to class and 0 if not) and $p(y)$ is the predicted probability that the record belongs to the class, for all N records. Along with that, we used a learning rate of 0.0005.

5. RESULTS

We evaluated our approach using the two datasets described in section 2.

5.1 First Dataset: ECG records

For the first dataset, *ECG records*, we will show the result for the 35 most relevant classes. Their results and the acronym by which they are referred in this paper can be seen in Tables I, II and III. Through this dataset we want to show how our method is able to output the correct result, and we compare our proposal to three baselines, described next.

Our first baseline is a regular expression (regex) classifier. Regular Expressions are a very simple technique. Some pattern is written in a formal language and given as input to the regex engine, which labels the text depending on the pattern being found in the text or not. This technique is used here to demonstrate the dataset complexity. We can see through Table II, in the Regex columns, that, even though the recall is equal to 1, which is expected in this baseline for this dataset, the precision is below 0.25 for all classes. This is also expected, since the input data is free text, with no clear pattern to be recovered using only regex.

The second baseline is a state-of-the-art topic modeling method, LDA [Yadav 2017; Allahyari et al. 2017]. LDA can be used as a classifier, in a non-supervised context, and works by connecting each document to each word by a thread, based on their appearance in the document and use this information to identify which documents discuss the same topic. Basically, it defines a relation between words and topics and determines the topics of a document, based on its words. This model was fine-tuned to our database to get the best results, although, at the end, our proposed method achieves the best result in all cases.

The third baseline is a model implemented with a transformer network architecture, based solely on attention mechanisms [Yang et al. 2016], dispensing recurrence and convolutions entirely. Experiments show these models to be superior in quality while being more parallelizable and requiring significantly less time to train [Vaswani et al. 2017]. Some tests have been conducted in order to define what would

Table I. Precision rates for four methods applied in the first test dataset. The best result amongst all classes is highlighted.

Acronym	Class	Precision				
		PM	**TF**	**LDA**	**Regex**	#
AI	Analysis Impossible due to Absence of Electrocardiographic Signal	**1.000**	**1.000**	0.090	0.023	30
LPFB	Left Posterior Fascicular Block	**0.967**	0.941	0.604	0.037	49
WPW	Wolff Parkinson White	**0.967**	0.909	0.857	0.023	31
LAFB	Left Anterior Fascicular Block	**0.964**	0.925	0.242	0.176	230
PMKR	Pacemaker	0.937	**0.967**	0.125	0.049	64
CDRB	Conduction Disorder of the Right Branch	0.920	**0.921**	0.871	0.047	61
PRWP	Poor R-wave Progression	**0.893**	0.853	0.809	0.046	61
RAD	Right Axis Deviation	0.891	**0.969**	0.305	0.053	69
PQTI	Prolonged QT Interval	**1.000**	0.893	**1.000**	0.026	34
SA	Sinus Arrhythmia	0.871	**0.872**	0.059	0.027	35
EAR	Ectopic Atrial Rhythm	**0.903**	0.879	0.040	0.023	30
CDLB	Conduction Desorder of the Left Branch	**0.897**	0.892	0.057	0.029	38
RBBB	Right Bundle Branch Block	**0.861**	0.817	0.158	0.151	196
PIE	Possible Inversion of Electrodes	**0.900**	0.750	0.041	0.023	30
LBBB	Left Bundle Branch Block	**0.914**	0.844	0.105	0.077	100
AFL	Atrial Flutter	0.909	0.939	**0.944**	0.027	35
AF	Atrial Fibrillation	0.807	0.842	**0.846**	0.054	71
LAE	Left Atrial Enlargement	**0.893**	0.878	0.701	0.077	100
STA	Supraventricular Tachycardia	**0.914**	0.833	0.044	0.030	39
SPRI	Short PR Interval	**0.775**	0.738	0.059	0.024	32
SCVR	Secondary Changes in Ventricular Repolarization	**0.847**	0.814	0.617	0.156	204
MAT	Multifocal Atrial Tachycardia	**0.867**	0.667	0.750	0.012	16
AVB1	First-Degree Atrioventricular Block	0.756	**0.824**	0.477	0.094	123
PCVR	Primary Changes in Ventricular Repolarization	**0.920**	0.732	0.644	0.046	60
LAD	Left Axis Deviation	**0.900**	0.836	0.834	0.233	303
NCVR	Nonspecific Changes in Ventricular Repolarization	0.796	**0.851**	0.185	0.179	233
NECG	Normal ECG	**0.804**	0.725	0.613	0.060	79
VES	Ventricular Extrasystoles	0.732	**0.764**	0.552	0.102	133
EIA	Electrically Inactive Area	**0.793**	0.779	0.393	0.046	60
SB	Sinus Bradycardia	0.778	**0.867**	0.467	0.045	59
SI	Subendocardial Ischemia	**0.727**	0.364	0.019	0.013	18
AVB2M1	2nd Degree Atrioventricular Block Mobitz I	**0.884**	0.862	0.750	0.025	33
LVH	Left Ventricular Hypertrophy	**0.740**	0.707	0.504	0.051	67
SVES	Supraventricular Extrasystoles	**0.649**	0.635	0.059	0.049	64
ST	Sinus Tachycardia	**0.571**	0.415	0.019	0.018	24
Average Values		0.856	0.814	0.424	0.062	80.31

be the best configuration for a transformer based model in our task. After these tests, the model that showed the best result among all was chosen and is displayed in Table III. Although transformers present all these advantages over other architectures and even though the best possible version of the technique for our case is manually chosen, our proposed model still performs better in most cases, outperforming the baseline techniques in most classes. These results show the efficiency of our method when compared to other models.

In the tables, we refer to PM as the **P**roposed **M**odel, LDA as Latent Dirichlet Allocation, REG as the application of a simple Regex in order to find the terms of the Diagnoses Dictionary in the reports and TF as the **T**rans**F**ormer based model.

We also present some Receiver Operator Characteristic (ROC) curves to help in the analysis of the technique performance. A ROC curve is a graph showing the performance of a classification model at different classification thresholds. This curve plots two parameters: the True Positive Rate and the False Positive Rate.

$$\textbf{True Positive Rate (TPR)} = \frac{TP}{TP+FN}$$

$$\textbf{False Positive Rate (FPR)} = \frac{FP}{TN+FP}$$

Table II. Recall rates for four methods applied in the first test dataset, ordered by F1. Here, although the best result is the Regex for all classes, we decided to highlight the best result amongst the three other classes. Regex always has a recall equal to 1 due to the form of the dataset and the nature of the method.

Acronym	Class	Recall				
		PM	**TF**	**LDA**	**Regex**	#
AI	Analysis Impossible due to Absence of Electrocardiographic Signal	0.966	**1.000**	1.0	1.000	30
LPFB	Left Posterior Fascicular Block	0.979	**0.980**	0.591	1.000	49
WPW	Wolff Parkinson White	0.967	**0.968**	0.967	1.000	31
LAFB	Left Anterior Fascicular Block	0.939	0.965	**0.995**	1.000	230
PMKR	Pacemaker	0.937	0.906	**1.000**	1.000	64
CDRB	Conduction Disorder of the Right Branch	0.950	**0.951**	0.557	1.000	61
PRWP	Poor R-wave Progression	**0.967**	0.951	0.557	1.000	61
RAD	Right Axis Deviation	**0.956**	0.913	0.782	1.000	69
PQTI	Prolonged QT Interval	**0.852**	0.735	0.558	1.000	34
SA	Sinus Arrhythmia	0.971	0.971	**1.000**	1.000	35
EAR	Ectopic Atrial Rhythm	0.933	0.967	**1.000**	1.000	30
CDLB	Conduction Desorder of the Left Branch	0.921	0.868	**0.973**	1.000	38
RBBB	Right Bundle Branch Block	0.954	0.980	**0.994**	1.000	196
PIE	Possible Inversion of Electrodes	0.900	0.900	**1.000**	1.000	30
LBBB	Left Bundle Branch Block	0.860	0.920	**0.980**	1.000	100
AFL	Atrial Flutter	0.857	0.886	**0.971**	1.000	35
AF	Atrial Fibrillation	**0.943**	0.901	0.929	1.000	71
LAE	Left Atrial Enlargement	0.800	0.860	**0.940**	1.000	100
STA	Supraventricular Tachycardia	0.820	0.897	**1.000**	1.000	39
SPRI	Short PR Interval	0.968	0.969	**1.000**	1.000	32
SCVR	Secondary Changes in Ventricular Repolarization	0.843	0.858	**0.887**	1.000	204
MAT	Multifocal Atrial Tachycardia	**0.812**	0.625	0.750	1.000	16
AVB1	First-Degree Atrioventricular Block	**0.935**	0.700	0.260	1.000	123
PCVR	Primary Changes in Ventricular Repolarization	0.766	**0.935**	0.816	1.000	60
LAD	Left Axis Deviation	0.778	**0.838**	0.429	1.000	303
NCVR	Nonspecific Changes in Ventricular Repolarization	0.875	0.906	**0.995**	1.000	233
NECG	Normal ECG	**0.835**	**0.835**	0.822	1.000	79
VES	Ventricular Extrasystoles	**0.924**	0.902	0.759	1.000	133
EIA	Electrically Inactive Area	0.833	0.883	**0.950**	1.000	60
SB	Sinus Bradycardia	0.830	**0.881**	0.830	1.000	59
SI	Subendocardial Ischemia	0.888	0.667	**1.000**	1.000	18
AVB2M1	2nd Degree Atrioventricular Block Mobitz I	0.696	**0.758**	0.545	1.000	33
LVH	Left Ventricular Hypertrophy	0.814	0.829	**0.895**	1.000	67
SVES	Supraventricular Extrasystoles	0.781	0.953	**1.000**	1.000	64
ST	Sinus Tachycardia	0.833	0.708	**1.000**	1.000	24
Average Values		0.884	**0.879**	0.850	1.000	80.31

Each point on the ROC curve represents a TPR vs. FPR pair corresponding to a particular decision threshold. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups [Fan et al. 2006]. An excellent model has AUC near to 1 which means it has a good measure of separability and when AUC is near 0.5, it means that the model has no class separation capacity whatsoever.

These graphs show very good results, e.g., among all curves, the one with the smallest AUC has an area of 0.94 and the best one has an area of 1.00. Two of the best curves can be seen in Figure 4 and Figure 5 and two of the worst curves can be seen in Figures 6 and 7. In the first example it can be seen that the model is able to achieve a perfect Sensitivity with a very small False Positive Rate. In the second example, it is possible to see how the worst curve is still a good result, showing an AUC of 0.94.

5.2 Second Dataset: Pacemaker patients

For the data set *Pacemaker patients* we display, in the Table IV, a subset of the results, showing the 10 classes of etiology for heart diseases, along with our method results. We display the same baselines

Table III. F1 rates for four methods applied in the first test dataset, ordered by F1. The last row indicates how many times the method has achieved the best F1 rate amongst all. In the case of a tie between models, all models with the top rate are considered best.

Acronym	Class	F1				
		PM	**TF**	**LDA**	**Regex**	#
AI	Analysis Impossible due to Absence of Electrocardiographic Signal	0.938	**1.000**	0.165	0.045	30
LPFB	Left Posterior Fascicular Block	**0.969**	0.960	0.597	0.072	49
WPW	Wolff Parkinson White	**0.967**	0.938	0.909	0.046	31
LAFB	Left Anterior Fascicular Block	**0.951**	0.945	0.390	0.300	230
PMKR	Pacemaker	**0.937**	0.935	0.222	0.094	64
CDRB	Conduction Disorder of the Right Branch	**0.935**	**0.935**	0.680	0.089	61
PRWP	Poor R-wave Progression	**0.929**	0.899	0.660	0.089	61
RAD	Right Axis Deviation	0.923	**0.940**	0.439	0.100	69
PQTI	Prolonged QT Interval	**0.920**	0.806	0.716	0.050	34
SA	Sinus Arrhythmia	0.918	**0.919**	0.111	0.052	35
EAR	Ectopic Atrial Rhythm	0.918	**0.921**	0.076	0.045	30
CDLB	Conduction Desorder of the Left Branch	**0.909**	0.880	0.107	0.056	38
RBBB	Right Bundle Branch Block	**0.905**	0.891	0.274	0.262	196
PIE	Possible Inversion of Electrodes	**0.900**	0.818	0.079	0.045	30
LBBB	Left Bundle Branch Block	**0.886**	0.880	0.190	0.143	100
AFL	Atrial Flutter	0.882	0.912	**0.957**	0.052	35
AF	Atrial Fibrillation	0.870	0.871	**0.885**	0.103	71
LAE	Left Atrial Enlargement	0.865	**0.869**	0.803	0.143	100
STA	Supraventricular Tachycardia	**0.864**	**0.864**	0.085	0.058	39
SPRI	Short PR Interval	**0.861**	0.838	0.113	0.048	32
SCVR	Secondary Changes in Ventricular Repolarization	**0.845**	0.835	0.728	0.27	204
MAT	Multifocal Atrial Tachycardia	**0.838**	0.645	0.750	0.024	16
AVB1	First-Degree Atrioventricular Block	**0.836**	0.757	0.336	0.172	123
PCVR	Primary Changes in Ventricular Repolarization	**0.836**	0.821	0.720	0.088	60
LAD	Left Axis Deviation	0.835	**0.837**	0.566	0.378	303
NCVR	Nonspecific Changes in Ventricular Repolarization	0.834	**0.877**	0.312	0.303	233
NECG	Normal ECG	**0.819**	0.776	0.702	0.114	79
VES	Ventricular Extrasystoles	0.817	**0.828**	0.639	0.185	133
EIA	Electrically Inactive Area	0.813	**0.828**	0.556	0.088	60
SB	Sinus Bradycardia	0.803	**0.874**	0.597	0.087	59
SI	Subendocardial Ischemia	**0.799**	0.471	0.037	0.027	18
AVB2M1	2nd Degree Atrioventricular Block Mobitz I	0.779	**0.806**	0.631	0.049	33
LVH	Left Ventricular Hypertrophy	**0.775**	0.763	0.645	0.098	67
SVES	Supraventricular Extrasystoles	0.709	**0.763**	0.112	0.093	64
ST	Sinus Tachycardia	**0.677**	0.523	0.038	0.036	24
Average Values		**0.866**	0.841	0.453	0.112	80.31
Best Models (count)		21	14	2	0	#

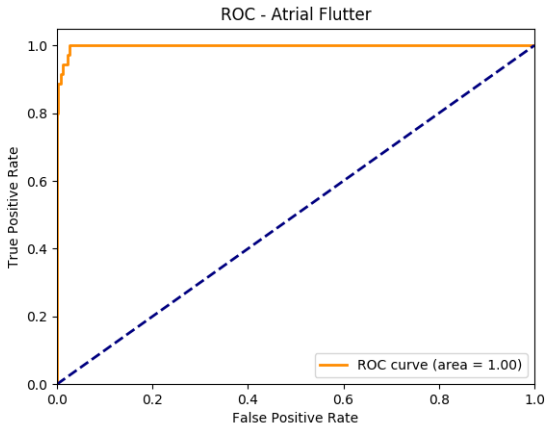

Fig. 4. ROC curve graph for the class "Atrial Flutter", an example of a very good result.

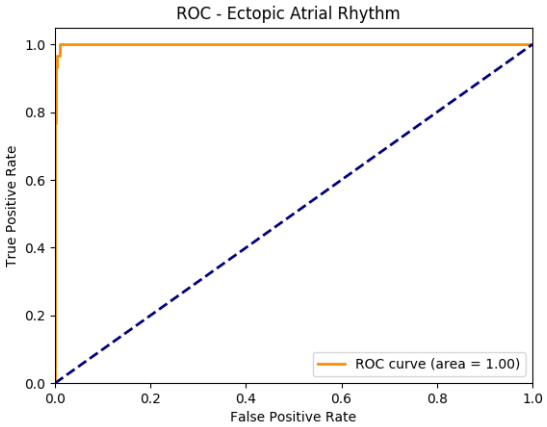

Fig. 5. ROC curve graph for the class "Ectopic Atrial Rhythm", another example of a good result.

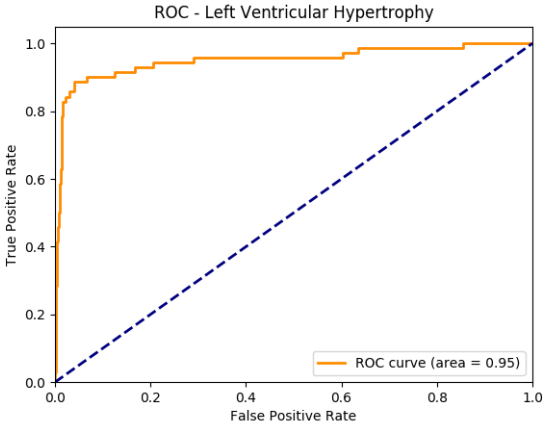

Fig. 6. ROC curve graph for the class "Left Ventricular Hypertrophy". One of the worst results.

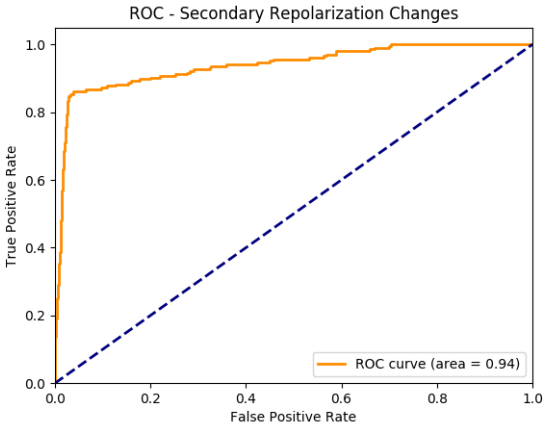

Fig. 7. ROC curve graph for the class "Secondary Repolarization Changes". The worst result amongst all curves.

Table IV. Precision, recall and F1 rates for the second application.

Class	Precision				Recall				F1			
	PM	TF	LDA	REG	PM	TF	LDA	REG	PM	TF	LDA	REG
Chagas	1.000	0.867	0.157	1.000	0.762	0.620	1.000	0.904	0.865	0.722	0.271	0.950
Schemic Cardiomyopathy	1.000	0.236	0.194	0.944	0.423	0.500	1.000	0.654	0.594	0.321	0.325	0.773
Valvular Heart Disease	1.000	0.244	0.185	1.000	0.600	0.880	1.000	0.400	0.750	0.383	0.312	0.571
Hypertrophic Cardiomyopathy	1.000	0.154	0.081	0.000	1.000	0.182	1.000	0.000	1.000	0.167	0.151	0.000
Congenic Cardiopatics	1.000	0.429	0.111	0.000	1.000	0.600	1.000	0.000	1.000	0.500	0.200	0.000
Long QT Syndrome	1.000	0.271	0.163	1.000	0.682	0.727	1.000	0.591	0.811	0.396	0.280	0.743
Brugada Syndrome	1.000	0.211	0.148	1.000	0.700	0.950	1.000	0.650	0.824	0.345	0.258	0.787
Idiopathic Ventricular Fibrillation	1.000	0.429	0.096	0.000	1.000	0.461	1.000	0.000	1.000	0.444	0.176	0.000
Arrhytthmogenic Dysplasia of VD	1.000	0.450	0.115	0.000	1.000	0.600	1.000	0.000	1.000	0.514	0.205	0.000
Idiopathic Cardiomyopathy	1.000	0.385	0.082	0.000	1.000	0.909	1.000	0.000	1.000	0.541	0.151	0.000

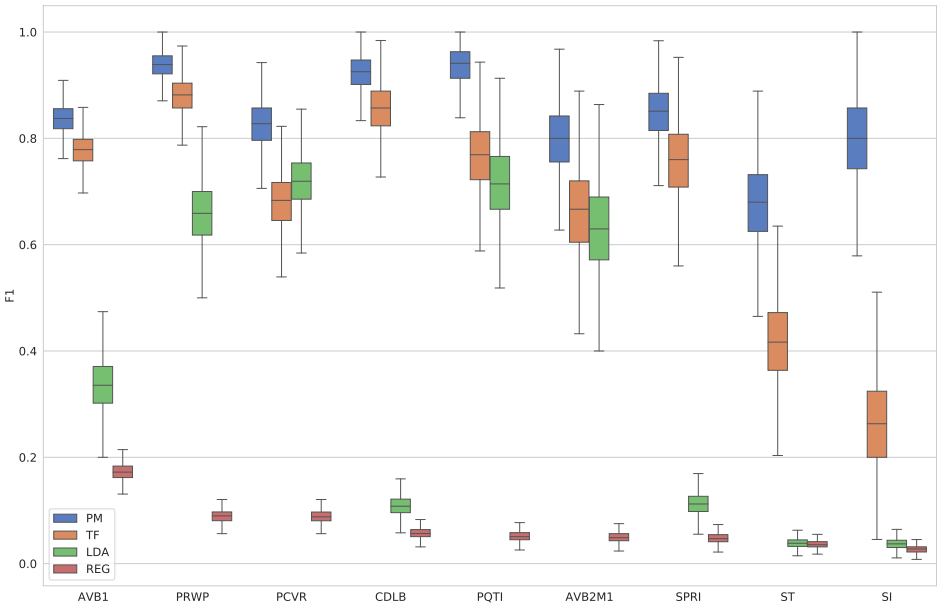

Fig. 8. Boxplot with F1 rates of each method on the Bootstrap samples. Here we show the classes in which the Proposed Method has achieved a statistically better performance.

used for the first application as comparison.

In this application, it can be seen that the baselines did not perform well, as a consequence of the database being much more complex and having a very wide range of terms for each class, many of which absent in our dictionary. Regex doesn't always achieve the maximum Recall (1) because sometimes the physicians specify that a report doesn't belong to a given class. Also, the fact that Regex obtained F1 equal to 0 in some classes, shows that, in these cases, none of the terms were written exactly like the terms in the Diagnoses Dictionary. However, even with these obstacles, our method was able to accomplish what none of the baselines could and achieved a very good performance in this database, achieving an F1 score above 0.8 in most classes.

5.3 Statistical Analysis

In order to improve the robustness of our results, we analyzed their statistical significance. We used bootstrapping to generate 5.000 samples of the test datasets, each with the same size of the original dataset, i.e., 1301 records for the First Dataset and 137 records for the Second Dataset. We performed the proposed method, as well as the three baselines, on all samples. The same metrics presented in the previous tables have been calculated for each class in each one of these samples.

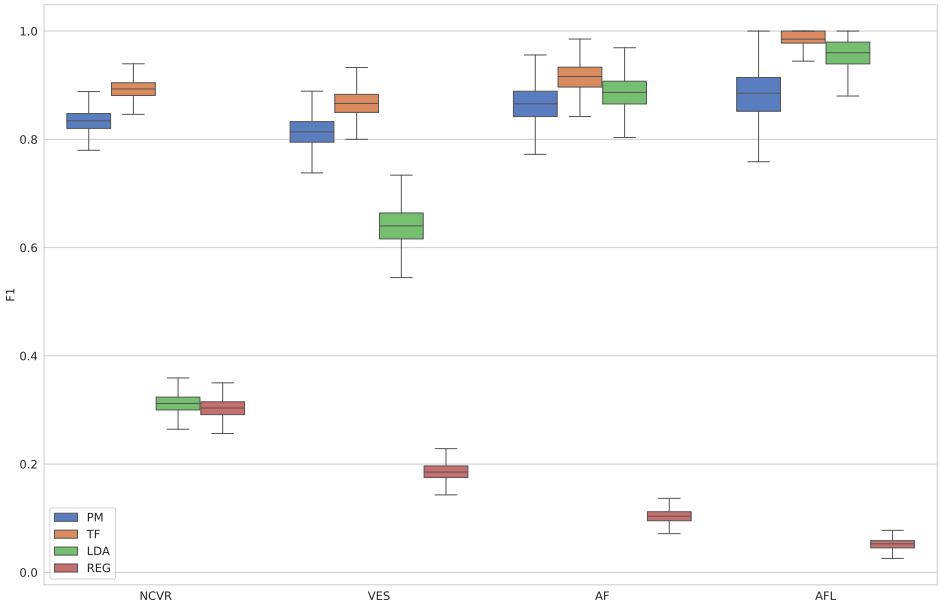

Fig. 9. Boxplot with F1 rates of each method on the Bootstrap samples. Here we show the classes in which the Proposed Method has achieved a statistically worse performance than some other method.

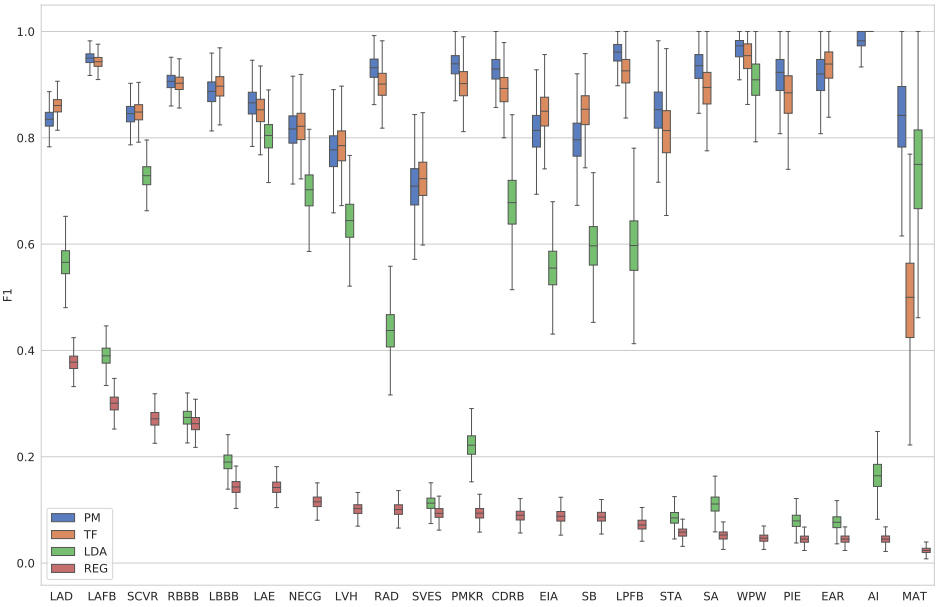

Fig. 10. Boxplot with F1 rates of each method on the Bootstrap samples. Here we show the classes in which the Proposed Method has achieved a performance that is at least statistically similar than any other method.

In Figures 8, 9 and 10, boxplots representing the performance of each method in each class for the first dataset can be seen. We divide the results into three graphs for a better visualization of the results. In Figure 8, we present the 9 classes in which the Proposed Method has achieved a statistically better performance than any other method, while in Figure 9 we present the 4 classes in which the Proposed Method has achieved a statistically worse performance than some other method, and, in Figure 10 we present the classes in which the Proposed Method has achieved a performance that is at

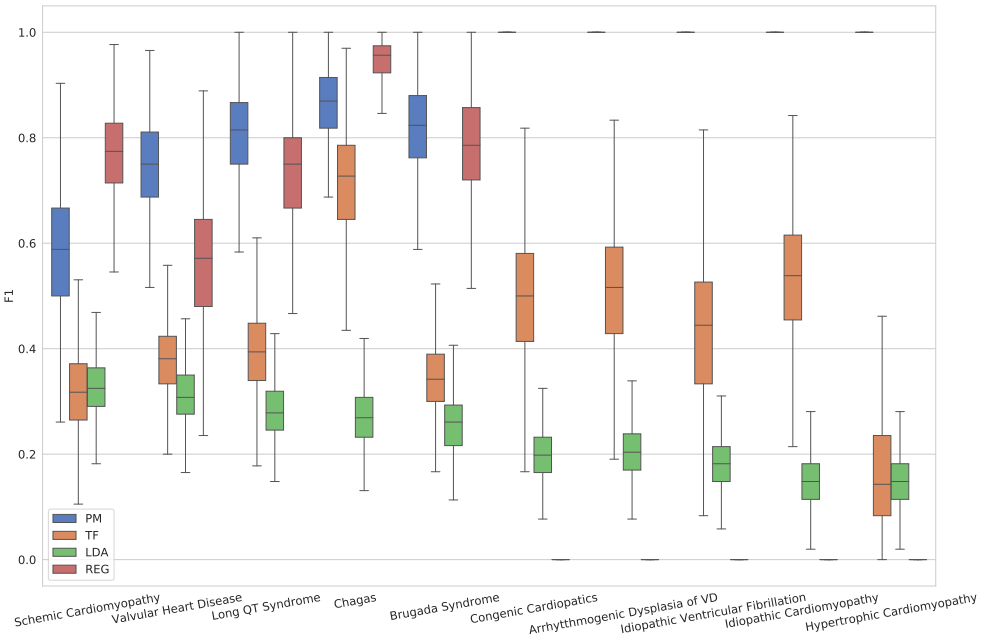

Fig. 11. Boxplot representing how well each method performed on each class on the bootstrap samples from the Second Dataset.

least statistically similar to any other method. In all cases of Figure 9, the method that has achieved the best performance is TF. It is important to point that Transformers are not a trivial baseline, as explained in section 3. They are the state of the art in text classification tasks and, even so, they only managed to overcome our method in 4 of the 35 classes. The fact that our method is superior to this baseline in 9 classes, while still managing to achieve a similar performance in the vast majority of the other classes, demonstrates its effectiveness.

In Figure 11, it is possible to see a boxplot representing the performance of each method in each class in the Bootstrap samples of the second test dataset. In this dataset, our method achieved a statistically superior performance in 6 cases. In 2 classes, the best method was the Regex and in 2 other classes there was no method that is statistically superior to its counterparts. The performance of the Regex is related to the fact that the classes on this dataset behave very differently. In the 6 cases where regex is better, it is because the classes are very well behaved, i.e., in these classes, it is common for physicians to write the diagnosis exactly as presented in the Diagnoses Dictionary, so a simple regular expression is enough to classify most records. In the other 6 cases, the opposite happens, as these are not well behaved classes and have their descriptions written in different ways by physicians in the reports. A possible reason why TF do not perform well on this dataset is because the amount of training data is much smaller, and this method requires a large database to be able to achieve a good performance. The fact that our method achieves good performance, even on a database with less amount of data for training, shows its robustness.

6. CONCLUSION

In this work we proposed and evaluated a method to map medical reports written in free text into labels that automated classifiers may use as input. Our method was applied to two real cardiology-related datasets and achieved good results in both, even when other techniques were not able to handle the complexity of the reports. We believe that even better results may be achieved using a more detailed Diagnoses Dictionary.

Several works support the development of techniques like ours as relevant [Prince and Roche 2009; Gabrieli and Speth 1990; Baud et al. 1992] and studies have demonstrated the need to apply techniques such as the one employed in this paper so that data can be used in an effective way [Ribeiro et al. 2020; Paixao et al. 2018; Hughes et al. 2004]. The results achieved demonstrate not only that our work is relevant, but also that it is applicable to a wide range of scenarios.

Finally, for future work, we intend to apply this technique in other scenarios and contexts to demonstrate the generality and assess the robustness of our method even further.

REFERENCES

Alkmim, M. B., Figueira, R. M., Marcolino, M. S., Cardoso, C. S., Abreu, M. P. d., Cunha, L. R., Cunha, D. F. d., Antunes, A. P., Resende, A. G. d. A., Resende, E. S., et al. Improving patient access to specialized health care: the telehealth network of minas gerais, brazil. *Bulletin of the World Health Organization* vol. 90, pp. 373–378, 2012.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Baud, R., Rassinoux, A.-M., and Scherrer, J.-R. Natural language processing and semantical representation of medical texts. *Methods of information in medicine* 31 (02): 117–125, 1992.

Collins, M. Three generative, lexicalised models for statistical parsing. *arXiv preprint cmp-lg/9706022*, 1997.

Dang PA, Kalra MK, B. M. e. a. Natural language processing using online analytic processing for assessing recommendations in radiology reports. *J Am Coll Radiol* vol. 5,3, pp. 197-204, 2008.

Fan, J., Upadhye, S., and Worster, A. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine* 8 (1): 19–20, 2006.

Ford, E., Nicholson, A., Koeling, R., Tate, A. R., Carroll, J., Axelrod, L., Smith, H. E., Rait, G., Davies, K. A., Petersen, I., et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? *BMC medical research methodology* 13 (1): 105, 2013.

Friedman, C. Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association, pp. 595, 1997.

Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S. B., and Clayton, P. D. Natural language processing in an operational clinical information system. *Natural Language Engineering* 1 (1): 83–108, 1995.

Gabrieli, E. R. and Speth, D. J. Automated analysis of medical text i. clue gathering. *Journal of medical systems* 14 (1-2): 71–91, 1990.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Harris, Z. S. Distributional structure. *Word* 10 (2-3): 146–162, 1954.

Hassanpour, S. and Langlotz, C. P. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine* vol. 66, pp. 29–39, 2016.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation* 9 (8): 1735–1780, 1997.

Hripcsak, G., Friedman, C., Alderson, P. O., DuMouchel, W., Johnson, S. B., and Clayton, P. D. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of internal medicine* 122 (9): 681–688, 1995.

Hu, R. and Singh, A. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.

Hughes, N. P., Tarassenko, L., and Roberts, S. J. Markov models for automated ecg interval analysis. In *Advances in Neural Information Processing Systems*. pp. 611–618, 2004.

Jagannatha, A. N. and Yu, H. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*. Vol. 2016. NIH Public Access, pp. 856, 2016.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 449–456, 2019.

Klein, D. and Manning, C. D. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*. pp. 423–430, 2003.

Mamlin, B. W., Heinze, D. T., and McDonald, C. J. Automated extraction and normalization of findings from cancer-related free-text radiology reports. In *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association, pp. 420, 2003.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pp. 3111–3119, 2013.

Paixao, G., Silva e Silva, L. G., Gomes, P., Ferreira, M., Oliveira, D., Ribeiro, M., Ribeiro, A., Nascimento, J., Cardoso, G., Araujo, R., et al. Clinical outcomes in digital electrocardiography: Evaluation of mortality in atrial fibrillation (code study). *Circulation* 138 (Suppl_1): A16594–A16594, 2018.

Pedrosa, J. A. O., Oliveira, D., Meira Jr, W., and Ribeiro, A. Automated classification of cardiology diagnoses based on textual medical reports. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, pp. 185–192, 2020.

Pradhan, N., Gyanchandani, M., and Wadhvani, R. A review on text similarity technique used in ir and its application. *International Journal of Computer Applications* 120 (9), 2015.

Prince, V. and Roche, M. *Information retrieval in biomedicine: natural language processing for knowledge integration*. Medical Information Science Reference New York, 2009.

Ramos, J. et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*. Vol. 242. Citeseer, pp. 29–48, 2003.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira Jr., W., Schön, T. B., and Ribeiro, A. L. P. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications* 11 (1): 1760, 2020.

Souza, R. C., de Brito, D. E., Cardoso, R. L., de Oliveira, D. M., Meira, W., and Pappa, G. L. An evolutionary methodology for handling data scarcity and noise in monitoring real events from social media data. In *Ibero-American Conference on Artificial Intelligence*. Springer, pp. 295–306, 2014.

Spyns, P. Natural language processing in medicine: an overview. *Methods of information in medicine* 35 (04/05): 285–301, 1996.

Stein HD, Nadkarni P, E. J. M. P. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository, 2000.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*. pp. 5998–6008, 2017.

Xu, J. and Sharma, P. Structured report data from a medical text report, 2019. US Patent App. 16/382,358.

Yadav, P. Patient report retrieval using semantic lda with cosine similarity. *Int. J. Innov. Sci. Eng. Technol.* 4 (7): 402–408, 2017.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. pp. 1480–1489, 2016.