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 diagnoses of diseases 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 disease 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.

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Information extraction**; • **Applied computing**;

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 category taxonomy. A key information that is usually present in such medical reports is the set of symptoms and possible disease diagnoses. But such information may be still limited w.r.t. disease categories and may not allow for the expression of nuances [Stein HD 2000]. As a consequence, free text analysis is commonly chosen as strategy when no category precisely describes clinical findings, or when there is a need to give supporting evidence for a 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 too time consuming to be achievable in most applications [Paixao et al. 2018], justifying the need for an automated solution.

This problem can 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]. Its effectiveness has been proven previously. Most implemented medical NLP systems reach an 80 - 85% range of recall and a precision of 95 - 99% [Mamlin et al. 2003]. Even though this is not a perfect result, it may be good enough for it to be used in real-world applications, since humans fall within the same performance range.

While there are several models that make use of *NLP* to retrieve and use reports' information [Fried-

man 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 classification model that is specifically developed for clinical reports written in Portuguese. The method map diagnosis labels to each textual report using only a dictionary that describes common terms for each diagnosis class. Because of that, no manual categorization of any textual report is necessary for the method to work. We explored the most recent developments in terms of embedding models to create a robust and efficient method and applied our model to two extremely unbalanced real cardiology datasets provided by Hospital das Clinicas de Minas Gerais, one of which comprises more than 2,000,000 reports.

2. RELATED WORKS

Several works emerged recently to automatically classify textual data. Despite the large volume of data associated with healthcare applications, a significant portion of these data is free text, without a clear pattern that an automated method can use as input. The usage of *NLP* for such problems was proposed by several works [Spyns 1996; Friedman 1997; Xu and Sharma 2019; Hassanpour and Langlotz 2016] in recent years, but it is still a big challenge.

Classic text extraction and model building requires a large database, manually labeled, to support some supervised learning algorithm to classify unlabeled reports [Souza et al. 2014; Jagannatha and Yu 2016]. However, it is not always possible to build a training dataset, and thus these data is not available for most applications.

Although there are works that use a semantic approach [Friedman et al. 1995; Spyns 1996] to retrieve information using characteristics of the language to improve the result, most of these works are designed for the English language, which makes it impossible to apply them to other languages.

We propose here a disease classification method for medical reports written in Portuguese, which differs from other approaches for not using manually labeled data. We evaluated our method in two real datasets.

There are three other methods that can be used to try to solve this same problem and that we will use as a base of comparison for our results: Regular Expressions, Latent Dirichlet Allocation (LDA) and Transformer Models. [Vaswani et al. 2017].

Some works [Yadav 2017; Allahyari et al. 2017] used a modified version of the Latent Dirichlet Allocation, using semantic terms to improve the result. The classical version of the LDA is a topic modeling in a non-supervised context, since LDA works by connecting each document to each word by a thread based on their appearance in the document and then use this information to know which documents discuss the same topic. Even though, there are works extended it towards a self-supervised version.

Transformer models, however, were developed to solve the problem of sequence transduction, or neural machine translation and are based solely on attention mechanisms [Bahdanau et al. 2014], a way to search for parts of a source sentence relevant to predicting a target word in an encoder-decoder model. That means any task that transforms an input sequence to an output sequence. This includes speech recognition, text-to-speech transformation andom labeling tasks.

3. METHODOLOGY

Our model is a self-supervised learning that learns the correct patterns from the target classes to perform the classification. In order to accomplish this task, we based our method in an embedding model that stacks three different classification steps.

It means that, for our method, it is not necessary a manually classified dataset and we divided the problem into three components. A flowchart of the process can be seen in Figure 1.

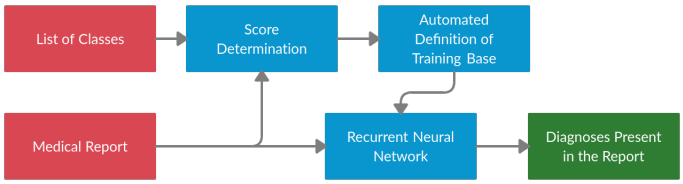

Fig. 1. Flowchart of the proposed approach. Here, the red rectangles represent the information that is used as input to the method, the blue rectangles represent the method steps and the green rectangles represent its output.

(1) For a fraction of the dataset, we determine a score for each text in relation to each class, based on the distance between words, using the classical Levenshtein string distance [Levenshtein 1966];
(2) We then make use of the known prevalence of each class, that is, the proportion of existing cases in a given population, to define what is the score threshold for it. We can then use this threshold to estimate, for each text, its class. Using this information, it is possible to automatically create a training database for a recurrent neural network;
(3) Finally, we instantiate a recurrent neural network and use the database that has been defined in the previous two steps to train it.

3.1 Score Determination

We built a dictionary that contains the common terms for each diagnosis class written as sentences, and we consider that it may contain just acronyms of the terms. The determination of the score is different for acronyms. If the acronym is a substring of the text report, then the result of the comparison is equal to 1, otherwise, 0. For usual terms, the comparison is defined as follows:

We denote $lev(A, B)$ as the Levenshtein string distance between strings A and B. Also, lets define a 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 this term of the class will be equal to:

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

The score between a report and a class is the maximum score for all terms of the class present in the dictionary.

3.2 Automated Definition of Training Base

Using the result of the score from the last step, we generate a training base for a sub process that will make use of supervised learning. This training base is automatically generated by our method and is a subset of the complete database. The subset consists of the records that had the highest score and is built in a way that the number of records belonging to each class is as close as possible to the real prevalence of the diagnoses. The optimal subset division value is called threshold, and depicted in Figure 2.

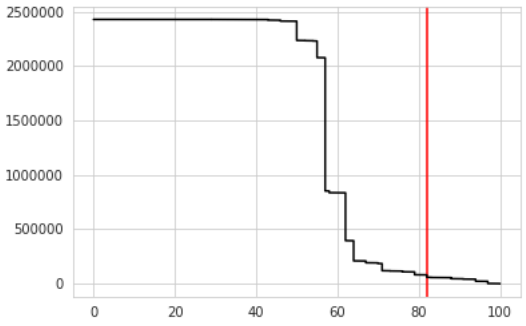

Fig. 2. Visual explanation of how a threshold is chosen. Axis Y represents number of registers that have a score greater or equal to the score defined in Axis X. The red line shows where the threshold must be placed for the number of registers to be as close as possible to the one defined by the class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

We can now use the training base to feed a machine learning algorithm. In our case we implemented a recurrent neural network to learn the patterns. Our hypothesis is that the subset created has latent features so that a supervised learning method can learn from them and then classify the entire dataset.

3.3 Application of the Recurrent Neural Network

The first step is to transform the medical reports into vectors that can be used as input for a neural network. In order to vectorize the reports we first created a dictionary, associating each word to an unique key. Since the number of different words in the reports can be massive, only the most common words were taken into account and the index of each word was related to its frequency in the reports, so lower integer means more frequent word. With this dictionary defined, we transformed each report into a vector, replacing each of its words with their key in the dictionary. Then the process of creating the neural network begins. It's structure is defined as follows:

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 that 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 by Hochreiter and Schmidhuber 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. Its results are presented in the next section.

4. RESULTS

We evaluated our approach using two different cardiology-related datasets. In both cases, the only manually labeled exams are those in the test dataset.

-2*Acronym	-2*Class	Precision				Recall				F1				-2*#
		PM	TF	LDA	Regex	PM	TF	LDA	Regex	PM	TF	LDA	Regex	
AI	Analysis Impossible due to Absence of Electrocardiographic Signal	1.000	1.000	0.090	0.023	0.966	1.000	1.0	1.000	0.983	**1.000**	0.165	0.045	30
LPFB	Left Posterior Fascicular Block	0.967	0.941	0.604	0.037	0.979	0.980	0.591	1.000	**0.969**	0.960	0.597	0.072	49
WPW	Wolff Parkinson White	0.967	0.909	0.857	0.023	0.967	0.968	0.967	1.000	**0.967**	0.938	0.909	0.046	31
LAFB	Left Anterior Fascicular Block	0.964	0.925	0.242	0.176	0.939	0.965	0.995	1.000	**0.951**	0.945	0.390	0.300	230
PMKR	Pacemaker	0.937	0.967	0.125	0.049	0.937	0.906	1.000	1.000	**0.937**	0.935	0.222	0.094	64
CDRB	Conduction Disorder of the Right Branch	0.920	0.921	0.871	0.047	0.950	0.951	0.557	1.000	**0.935**	**0.935**	0.680	0.089	61
PRWP	Poor R-wave Progression	0.893	0.853	0.809	0.046	0.967	0.951	0.557	1.000	**0.929**	0.899	0.660	0.089	61
RAD	Right Axis Deviation	0.891	0.969	0.305	0.053	0.956	0.913	0.782	1.000	0.923	**0.940**	0.439	0.100	69
PQTI	Prolonged QT Interval	1.000	0.893	1.000	0.026	0.852	0.735	0.558	1.000	**0.920**	0.806	0.716	0.050	34
SA	Sinus Arrhythmia	0.871	0.872	0.059	0.027	0.971	0.971	1.000	1.000	0.918	**0.919**	0.111	0.052	35
EAR	Ectopic Atrial Rhythm	0.903	0.879	0.040	0.023	0.933	0.967	1.000	1.000	0.918	**0.921**	0.076	0.045	30
CDLB	Conduction Desorder of the Left Branch	0.897	0.892	0.057	0.029	0.921	0.868	0.973	1.000	**0.909**	0.880	0.107	0.056	38
RBBB	Right Bundle Branch Block	0.861	0.817	0.158	0.151	0.954	0.980	0.994	1.000	**0.905**	0.891	0.274	0.262	196
PIE	Possible Inversion of Electrodes	0.900	0.750	0.041	0.023	0.900	0.900	1.000	1.000	**0.900**	0.818	0.079	0.045	30
LBBB	Left Bundle Branch Block	0.914	0.844	0.105	0.077	0.860	0.920	0.980	1.000	**0.886**	0.880	0.190	0.143	100
AFL	Atrial Flutter	0.909	0.939	0.944	0.027	0.857	0.886	0.971	1.000	0.882	0.912	**0.957**	0.052	35
AF	Atrial Fibrillation	0.807	0.842	0.846	0.054	0.943	0.901	0.929	1.000	0.870	0.871	**0.885**	0.103	71
LAE	Left Atrial Enlargement	0.893	0.878	0.701	0.077	0.800	0.860	0.940	1.000	0.865	**0.869**	0.803	0.143	100
STA	Supraventricular Tachycardia	0.914	0.833	0.044	0.030	0.820	0.897	1.000	1.000	**0.864**	**0.864**	0.085	0.058	39
SPRI	Short PR Interval	0.775	0.738	0.059	0.024	0.968	0.969	1.000	1.000	**0.861**	0.838	0.113	0.048	32
SCVR	Secondary Changes in Ventricular Repolarization	0.847	0.814	0.617	0.156	0.843	0.858	0.887	1.000	**0.845**	0.835	0.728	0.27	204
MAT	Multifocal Atrial Tachycardia	0.867	0.667	0.750	0.012	0.812	0.625	0.750	1.000	**0.838**	0.645	0.750	0.024	16
AVB1	First-Degree Atrioventricular Block	0.756	0.824	0.477	0.094	0.935	0.700	0.260	1.000	**0.836**	0.757	0.336	0.172	123
PCVR	Primary Changes in Ventricular Repolarization	0.920	0.732	0.644	0.046	0.766	0.935	0.816	1.000	**0.836**	0.821	0.720	0.088	60
LAD	Left Axis Deviation	0.900	0.836	0.834	0.233	0.778	0.838	0.429	1.000	0.835	**0.837**	0.566	0.378	303
NCVR	Nonspecific Changes in Ventricular Repolarization	0.796	0.851	0.185	0.179	0.875	0.906	0.995	1.000	0.834	**0.877**	0.312	0.303	233
NECG	Normal ECG	0.804	0.725	0.613	0.060	0.835	0.835	0.822	1.000	**0.819**	0.776	0.702	0.114	79
VES	Ventricular Extrasystoles	0.732	0.764	0.552	0.102	0.924	0.902	0.759	1.000	0.817	**0.828**	0.639	0.185	133
EIA	Electrically Inactive Area	0.793	0.779	0.393	0.046	0.833	0.883	0.950	1.000	0.813	**0.828**	0.556	0.088	60
SB	Sinus Bradycardia	0.778	0.867	0.467	0.045	0.830	0.881	0.830	1.000	0.803	**0.874**	0.597	0.087	59
SI	Subendocardial Ischemia	0.727	0.364	0.019	0.013	0.888	0.667	1.000	1.000	**0.799**	0.471	0.037	0.027	18
AVB2M1	2nd Degree Atrioventricular Block Mobitz I	0.884	0.862	0.750	0.025	0.696	0.758	0.545	1.000	0.779	**0.806**	0.631	0.049	33
LVH	Left Ventricular Hypertrophy	0.740	0.707	0.504	0.051	0.814	0.829	0.895	1.000	**0.775**	0.763	0.645	0.098	67
SVES	Supraventricular Extrasystoles	0.649	0.635	0.059	0.049	0.781	0.953	1.000	1.000	0.709	**0.763**	0.112	0.093	64
ST	Sinus Tachycardia	0.571	0.415	0.019	0.018	0.833	0.708	1.000	1.000	**0.677**	0.523	0.038	0.036	24
Average Values		0.856	0.814	0.424	0.062	0.884	0.879	0.850	1.000	0.866	0.841	0.453	0.112	80.31
Best Models (count)										21	14	2	0	#

Table I. Precision, recall and F1 rates in three methods applied of the 10 better and 10 worse results in the first test dataset, ordered by F1. Here PM is the **P**roposed **M**odel, LDA is Latent Dirichlet Allocation, REG is the application of a simple Regex in order to find the terms of the dictionary in the reports and TF is the **T**rans**F**ormer based model. In the last row the draws are counted twice

4.1 First Dataset: ECG records

The first dataset we used 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]. A dictionary containing common terms for 68 ECG abnormalities was used. For this work we will show the result for the 35 most relevant classes. Their results and the acronym by which they are referred as in this paper can be seen in Table I. Through this dataset we want to show how our method is able to give the correct result, and we compare our proposal to three baselines, described next.

Our first baseline is a regular expression (regex) that demonstrates the dataset complexity. We can see through Table I, in the Regex columns, that, even though the recall is equal to 1, which is expected in this baseline, the precision is below 0.25 for all classes. This is expected since the data is a free text, with no clear pattern to be recovered using only regex.

The second baseline is a state-of-the-art unsupervised text classifier, LDA [Yadav 2017; Allahyari et al. 2017]. This model was fine-tuned for our database to get the best results, but even so we can see that our proposed method gets 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 experiments have been conducted in order to define what would be the best configuration for a transformer based model in our task. Among all experiments, the model with the best result was chosen and is displayed in Table I. Even with transformers having all these advantages over other architectures and even though the best possible version of the technique for our case was chosen, our proposed model is still superior in most classes. These results show the efficiency of our method when compared to other models.

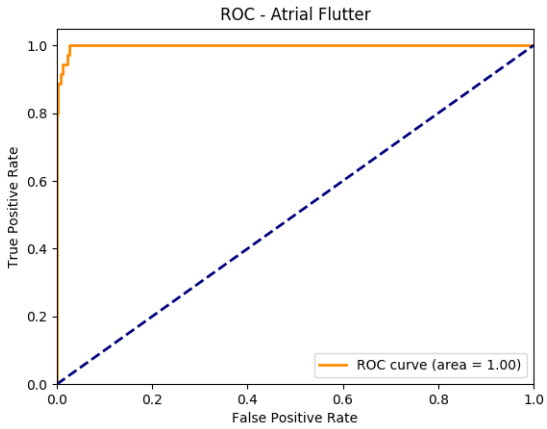

Fig. 3. ROC curve graph for the class "Atrial Flutter", an example of a very good result.

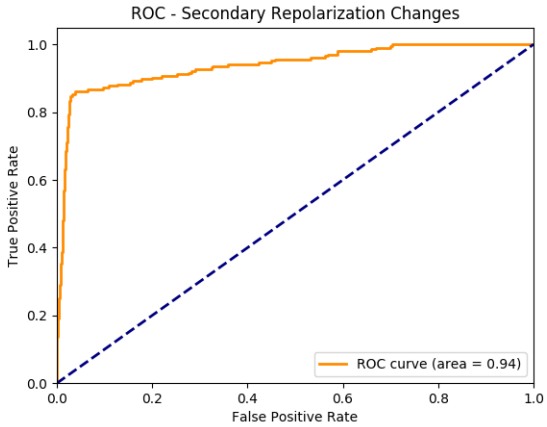

Fig. 4. ROC curve graph for the class "Secondary Repolarization Changes". The worst result amongst all curves.

Some Receiver Operator Characteristic (ROC) curves have been built to help in the analysis of the technique performance. In a ROC curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate (100 - Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal) [Fan et al. 2006]. Very good results can be seen in these graphs, e.g., among all curves, the one with the smallest AUC has an area of 0.94 and the one with the biggest has an area of 1.00. One of the best and the worst curve can be seen in Figure 3 and Figure 4. 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 amongst all is still a good result, showing an AUC of 0.94.

4.2 Second Dataset: Pacemaker patients

The second dataset contains records from pacemaker patients. The technique was also applied to this database to show the consistency of the model. Acquired through the Telehealth Network of Minas Gerais (TNMG), this dataset is compose of 70,312 records from 2,899 patients from Hospital das

Clinicas de Minas Gerais. We display a subset of the results, with the 10 most important classes of this second application, in the Table II. With this database, it is possible to demonstrate our method generality, regardless the context.

Class	Precision	Recall	F1
Chagas Disease	0.933	0.875	0.903
Schemic Cardiomyopathy	0.800	1.00	0.888
Valvular Heart Disease	1.000	1.000	1.000
Hypertrophic Cardiomyopathy	0.733	1.000	0.846
Congenic Cardiopatics	1.000	1.000	1.000
Long QT Syndrome	1.000	1.000	1.000
Brugada Syndrome	0.933	1.000	0.965
Idiopathic Ventricular Fibrillation	0.866	1.000	0.928
Arrhythmogenic Dysplasia do VD	1.000	1.000	1.000
Idiopathic Cardiomyopathy	0.733	1.000	0.846

Table II. Precision, recall and F1 rates of the model in the second application.

5. 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 can be achieved using a more detailed class dictionary.

Several works explain why the development of techniques like ours is so important [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]. This demonstrates not only that our work is relevant, but also that there is a large space for it's application.

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

6. ACKNOWLEDGEMENT

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.

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.

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.

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.

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.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*. Vol. 10. pp. 707–710, 1966.

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.

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

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.