NLP SUMMIT HEALTHCARE

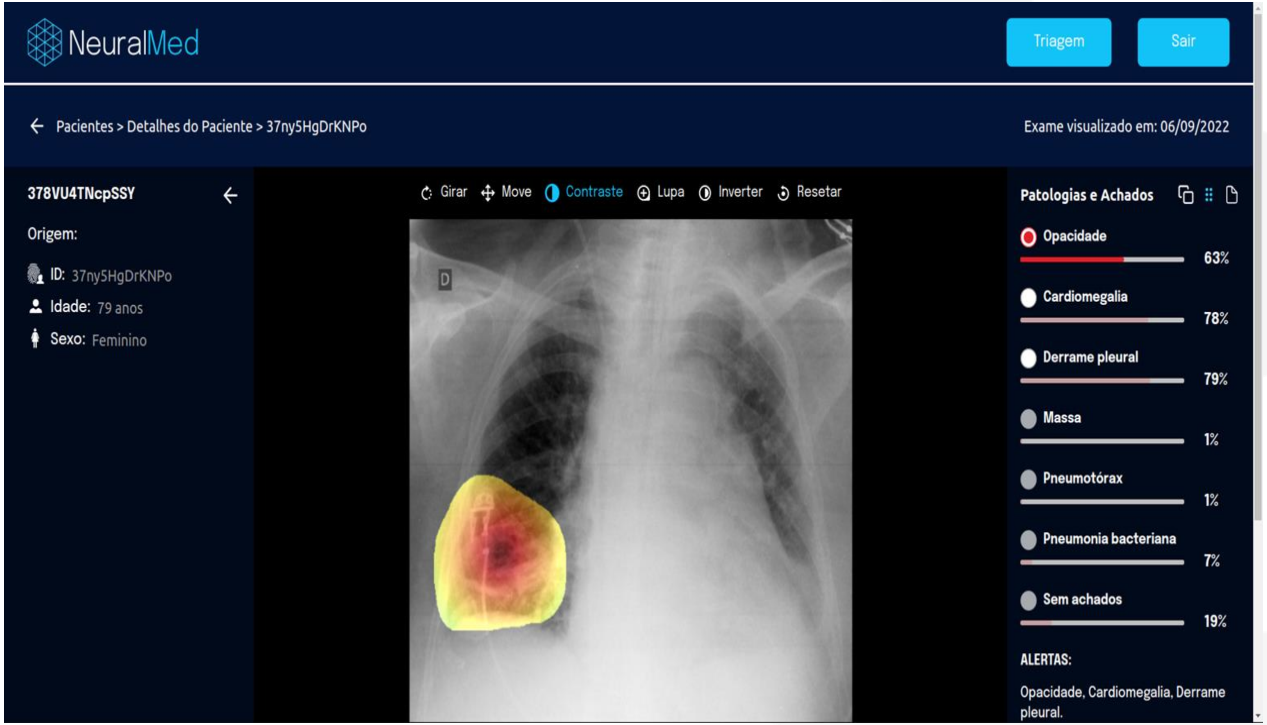
Using NLP to Structure Medical Records and Identify Patients with Abnormalities

Jéssica dos Santos

Head of Data
NeuralMed

NeuralMed

- Brazilian healthtech
- Started analyzing and classifying chest x ray exams
- Nowadays extracts and generate value from unstructured data, such as x-ray or ECG exams, medical reports and medical records.



Pacientes > Detalhes do Paciente > 37ny5HgDrKNPo

Exame visualizado em: 06/09/2022

378VU4TNcpSSY

Origem:

ID: 37ny5HgDrKNPo

Idade: 79 anos

Sexo: Feminino

Patologias e Achados

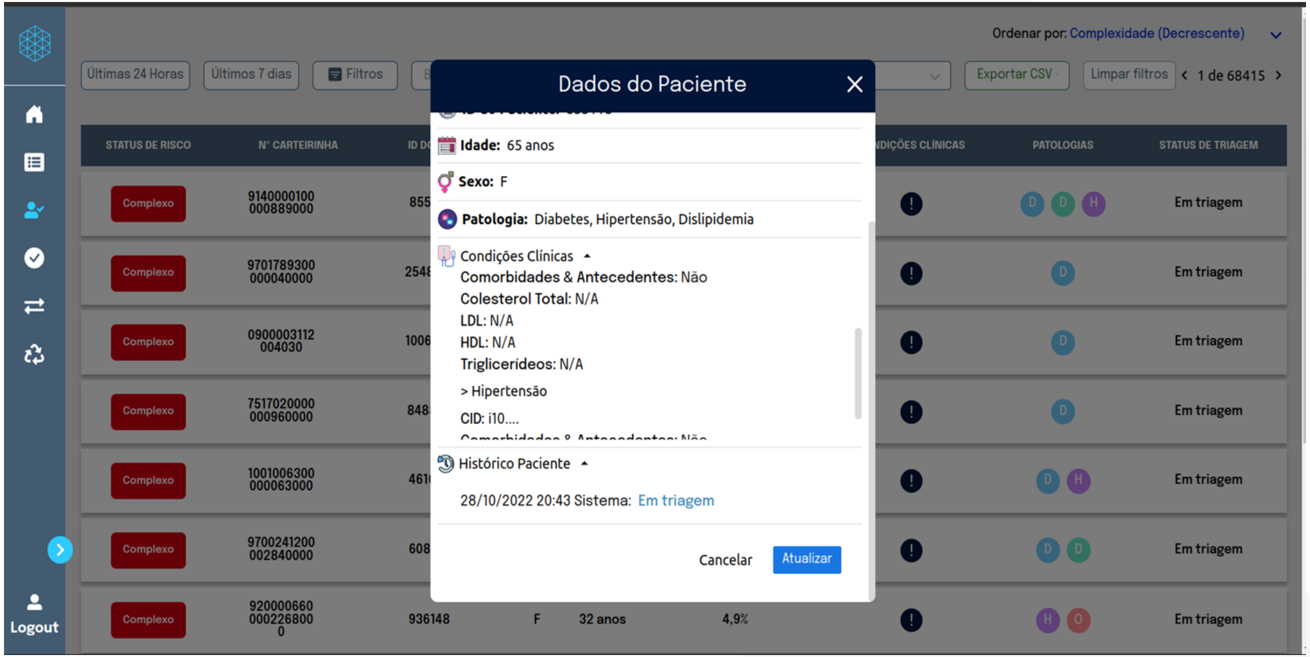
○ Opacidade	63%
● Cardiomegalia	78%
● Derrame pleural	79%
● Massa	1%
● Pneumotórax	1%
● Pneumonia bacteriana	7%
● Sem achados	19%

ALERTAS:

Opacidade, Cardiomegalia, Derrame pleural.

NeuralMed

- Brazilian healthtec
- Starts analyzing and classifying chest x ray exams
- Today extracts and generate value from unstructured data, such as x-ray or ECG exams, medical reports and medical records.



Ordenar por: Complexidade (Decrescente)

STATUS DE RISCO	N° CARTEIRINHA	ID DC				NDIÇÕES CLÍNICAS	PATOLOGIAS	STATUS DE TRIAGEM
Complexo	9140000100 000889000	855				!	D D H	Em triagem
Complexo	9701789300 000040000	2548				!	D	Em triagem
Complexo	0900003112 004030	1006				!	D	Em triagem
Complexo	7517020000 000960000	848				!	D	Em triagem
Complexo	1001006300 000063000	461				!	D H	Em triagem
Complexo	9700241200 002840000	608				!	D D	Em triagem
Complexo	920000660 000226800 0	936148	F	32 anos	4,9%	!	H O	Em triagem

Dados do Paciente

Idade: 65 anos

Sexo: F

Patologia: Diabetes, Hipertensão, Dislipidemia

Condições Clínicas

Comorbidades & Antecedentes: Não

Colesterol Total: N/A

LDL: N/A

HDL: N/A

Triglicerídeos: N/A

> Hipertensão

CID: i10....

Comorbidades & Antecedentes: Não

Histórico Paciente

28/10/2022 20:43 Sistema: Em triagem

NLP SUMMIT

Objective

Extract from medical records valuable information about patients, such as:

- **Family History** that can identify possible risks for diseases,
- **Medications** that patient takes or has already taken, which indicates the medical changes they had.
- **Exams results** that can diagnosis some condition
- History of **Comorbidities** that shows previous conditions
- Among other informations that can aid the identification of diseases from the patient

Problem

The information is written in free text in the most diverse ways, which makes it difficult for other physicians to use this information and even for hospital management.

Problem

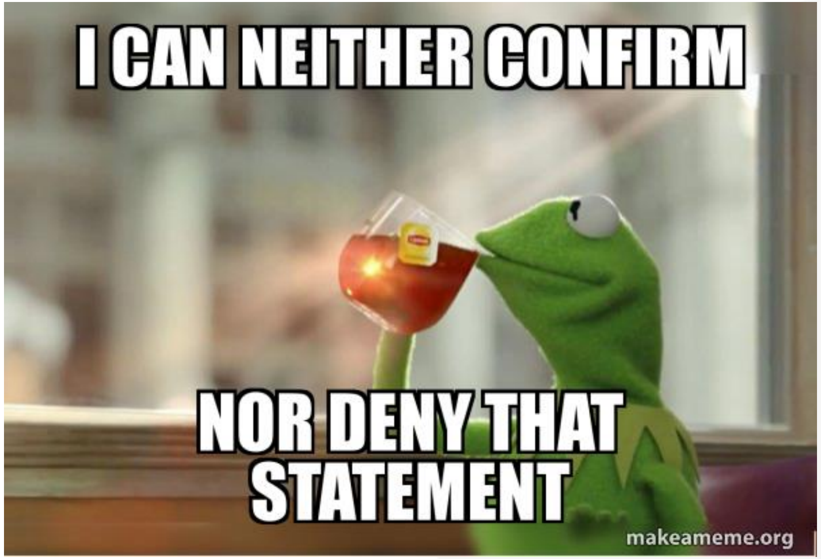
We always need to consider and understand what is being denied:

Nega alergias e DM, tem HAS
Denies allergies and DM, has HTN.

nega alergia medicamentosa e comorbidades como HAS, DM
Denies drug allergy and comorbidities such as HTN, DM

DM (-), HAS (+)
DM (-), HTN (+)

Filho de Mãe diabética, nega diabetes
Son of a diabetic mother, denies diabetes



Problem

The information is written in free text in the most diverse ways, which makes it difficult for other physicians to use this information and even for hospital management.

"Nega Alergias. Comorbidades: diabetes"
Deny Allergies. Comorbidities: diabetes

“Nega alergias, diabetes e hipertensão”
Deny Allergies, diabetes and, hypertension

“Nega alergias, diabetes. Hipertensão”
Deny Allergies, diabetes. Hypertension

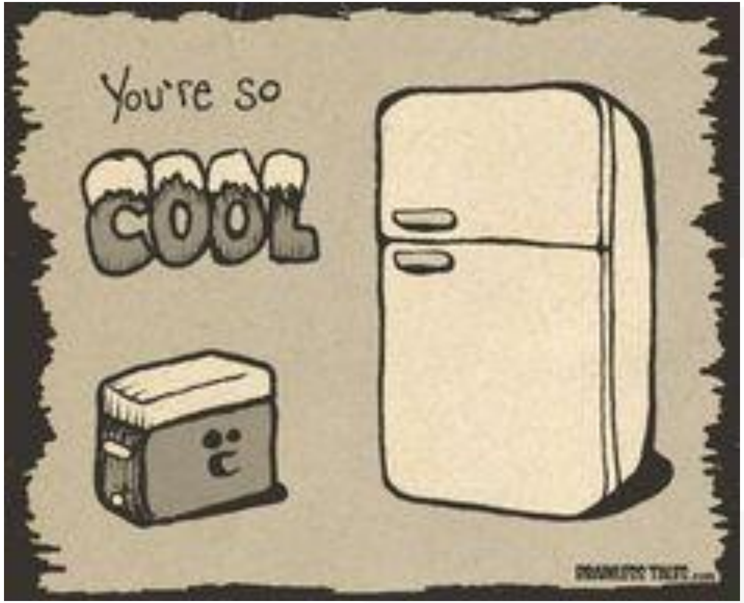
“* * * Alergias **alergia** não * * * Comorbidades **diabetes** não”
*** * * Allergies **allergy** no * * * Comorbidities **diabetes** no**

Problem

The same word not always means the same thing:

- COMORBIDADES COMO HAS E **DM** EM USO DE MEDICACAO

- FO com CTV oclusivo limpo e seco superficialmente, **DM** com debito hematico no frasco coletor, CTV em MID enfaixado limpo e seco superficialmente



In the first example DM is for diabetics, but in the second one is actually for Mediastinal Drainage

* didn't translate because the abbreviations will be different in english

Attemps

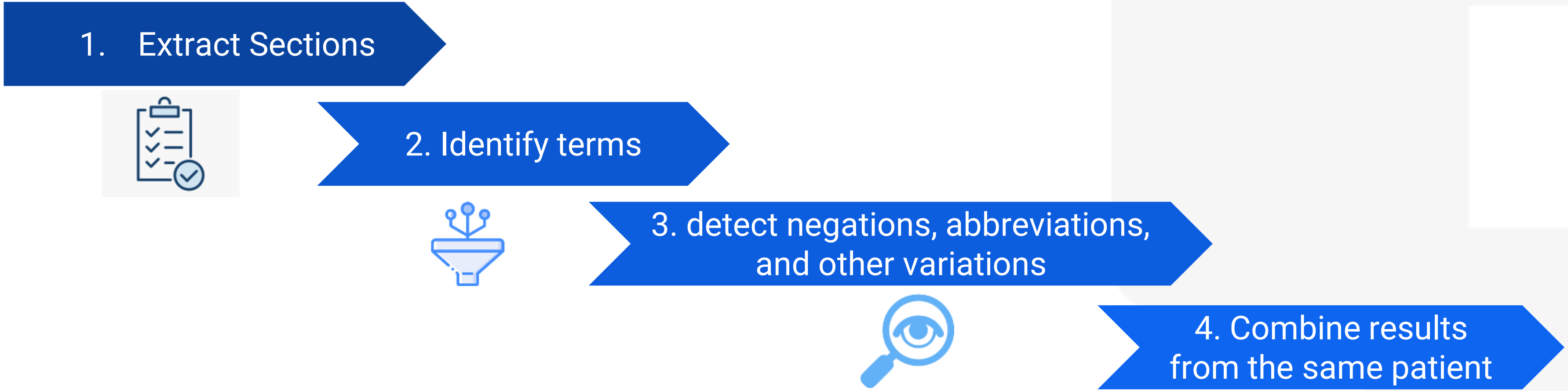
- Q&A

  BigBird based model with 2560 tokens finetuned with portuguese version of SQUAD dataset additionally with our own data with more than 500 questions of medical contexts extracted from EHR.

Solution

We used a set of NLP models combined with some data treatments and medical definitions to structure this information from the reports and use them to extract alterations such as diabetes and arterial hypertension.

Combining classical solutions with modern algorithms:



1. Extract Sections

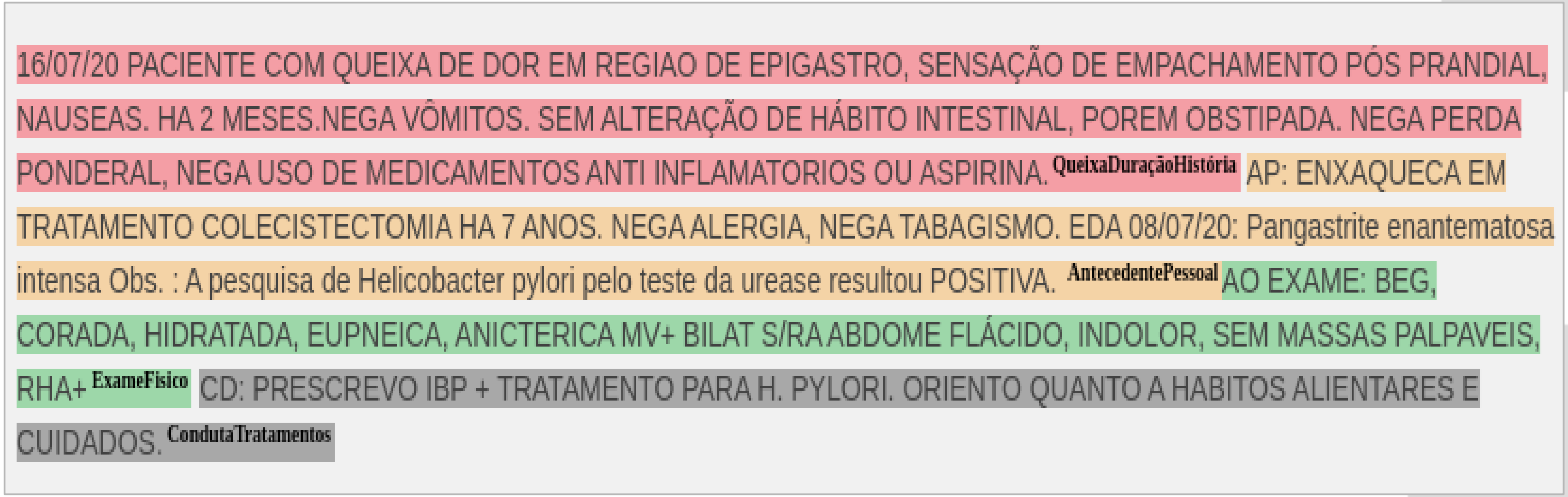
2. Identify terms

3. detect negations, abbreviations, and other variations

4. Combine results from the same patient

Solution

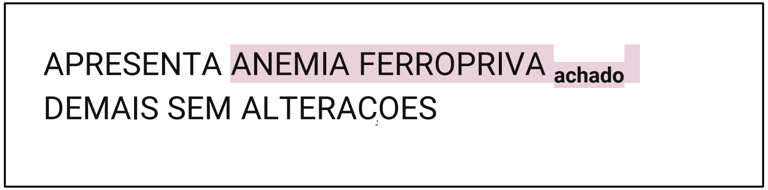
NER model to identify sections of the report.
Since we have a large amount of texts to process, over 6 millions, we prefer to use a simpler but faster model: CRF



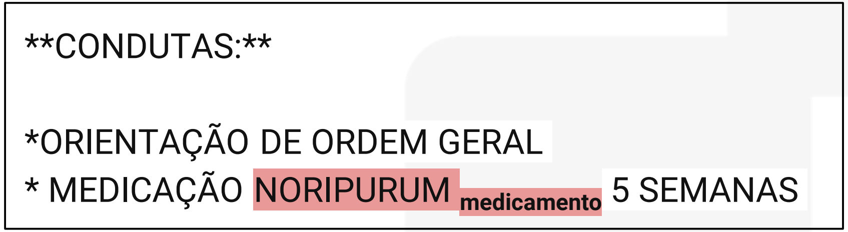
16/07/20 PACIENTE COM QUEIXA DE DOR EM REGIAO DE EPIGASTRO, SENSAÇÃO DE EMPACHAMENTO PÓS PRANDIAL, NAUSEAS. HA 2 MESES.NEGA VÔMITOS. SEM ALTERAÇÃO DE HÁBITO INTESTINAL, POREM OBSTIPADA. NEGA PERDA PONDERAL, NEGA USO DE MEDICAMENTOS ANTI INFLAMATORIOS OU ASPIRINA. QueixaDuraçãoHistória AP: ENXAQUECA EM TRATAMENTO COLECISTECTOMIA HA 7 ANOS. NEGA ALERGIA, NEGA TABAGISMO. EDA 08/07/20: Pangastrite enantematosa intensa Obs. : A pesquisa de Helicobacter pylori pelo teste da urease resultou POSITIVA. AntecedentePessoal AO EXAME: BEG, CORADA, HIDRATADA, EUPNEICA, ANICTERICA MV+ BILAT S/RA ABDOME FLÁCIDO, INDOLOR, SEM MASSAS PALPAVEIS, RHA+ ExameFisico CD: PRESCREVO IBP + TRATAMENTO PARA H. PYLORI. ORIENTO QUANTO A HABITOS ALIENTARES E CUIDADOS. CondutaTratamentos

Solution

Then, with each section splitted, we look for relevant information in that specific section. For example, in the section that describes Comorbidities, we will use another NER to identify specific diseases.

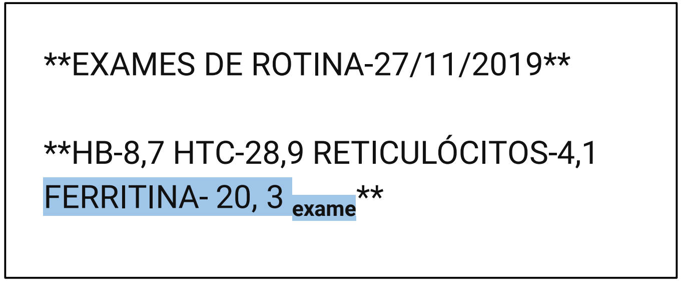


APRESENTA ANEMIA FERROPRIVA achado
DEMAIS SEM ALTERACOES



CONDUTAS:

*ORIENTAÇÃO DE ORDEM GERAL
* MEDICAÇÃO NORIPURUM medicamento 5 SEMANAS

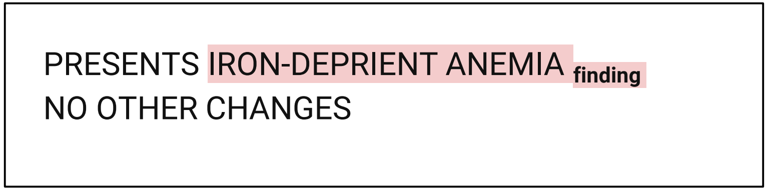


EXAMES DE ROTINA-27/11/2019

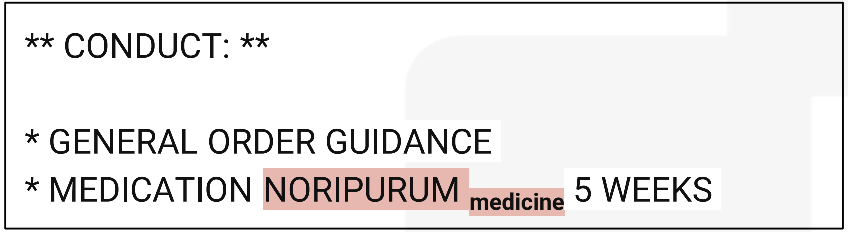
**HB-8,7 HTC-28,9 RETICULÓCITOS-4,1
FERRITINA- 20, 3 exame**

Solution

We trained a model based on BioBERTpt - Portuguese Clinical and Biomedical BERT, but with our own EHR dataset and annotations.

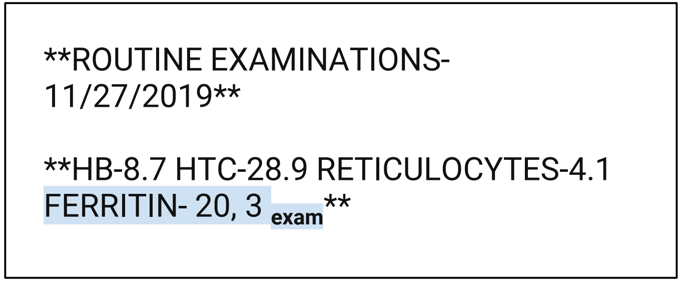


PRESENTS IRON-DEPRIENT ANEMIA finding
NO OTHER CHANGES



** CONDUCT: **

* GENERAL ORDER GUIDANCE
* MEDICATION NORIPURUM medicine 5 WEEKS



ROUTINE EXAMINATIONS-11/27/2019

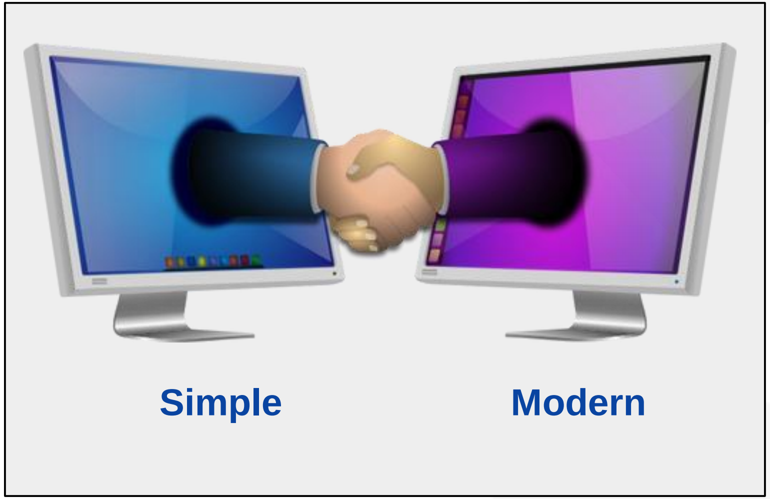
HB-8.7 HTC-28.9 RETICULOCYTES-4.1 FERRITIN- 20, 3 exam

Solution

Finally, to fix other problems such as: identify negations, abbreviations and missing of identification we create a combination of small but powerful specific solutions that uses:

- Genetic algorithm,
- Part of speech recognition and
- Regex.

Using classic algorithms with the others makes our solution powerful



Data

We have access to millions of reports, unfortunately annotation is expensive and, not all the reports presents all of the informations, so the models were trained with something about 20 thousands of samples.

The samples were annotated by our on time of medicals and students of medicine.

Results

- Our entire system is able to process 300k EHRs per day
- All of our models presents over 97% of accuracy
- In our first trial, with 300 patients identified with diabetics, 100% of them really have diabetics

Results

When we tested our solution in a big hospital in São Paulo we found about **71 thousands patients** with plurimetabolic syndrome, which is **5 times more** the amount of patients previously known.

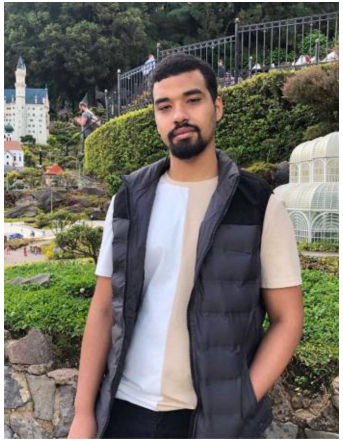
With the solution the hospital were able to add at least part of those patients in prevention programs, **improving patients' quality of life** and reducing hospital costs.

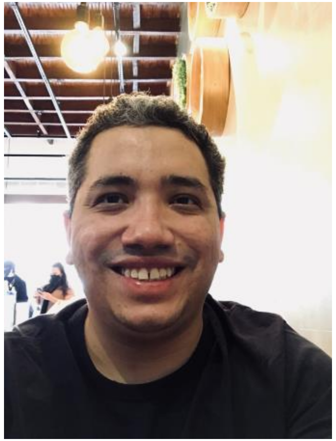
Future works

- Optimize process
- Decrease the number of models
- Try the new state of art in NLP
- Increase the number of diseases identified

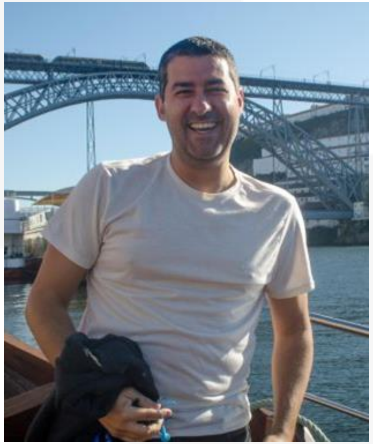
Acknowledgment

NLP Team





Medical Team



NLP SUMMIT HEALTHCARE

Presented by

John Snow LABS

www.nlpsummit.org