DEEPHEALTH

Hackaton - Course 2

De-identification of Radiological reports

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Index

1. Radiology Reports
    a. Electronic Health Records (EHR)
    b. Medical Texts
    c. Radiology Reports
2. De-identification methodologies
    a. Target Detection
    b. Regular Expressions and List Matching
    c. Named Entity Recognition
    d. De-identification
3. Named Entity Recognition - NLP
    a. NER - Machine Learning
    b. Corpus construction
    c. NER metrics
    d. Applications
4. Dismed
    a. The dataset
    b. Corpus
    c. Pipeline
    d. Results

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

DEEPHEALTH

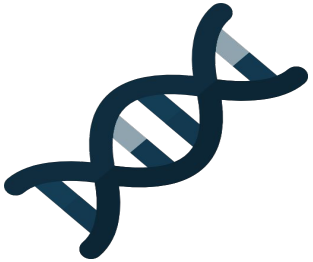
Radiology reports

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

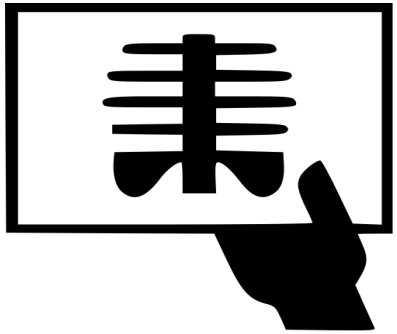
Electronic Health Records (EHR)

"Systematized collection of patient and population electronically stored health information in a digital format"
Structured and (mostly) Unstructured Data

Tests results



Omics Data



Medical Imaging

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Medical Texts

Clinical data in biomedical research

- Clinical Trial Protocols
- Physician Progress Notes
- Surgery Reports
- Radiology Reports

A 12-year old girl with known hyperagglutinability, presented to the emergency department with a 2-week history of headeaches and facial weakness. Neurologic examination indicated sensorineural hearing loss on the right side with Weber's test lateralizing to the left, and the Rinne's test demonstrating bone conduction greater than air conduction on the right. Magnetic resonance imaging of the head revealed severe structural defects of the right petrous temporal bone. No indication of cerebral infarction.

Figure 1, Starlinger, J. et al. (2016)

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Radiology Reports

Clinical data in biomedical research

0001 REPORT : REASON FOR EXAM (Entered by ordering clinician into CRIS): hx of head and neck cancer. needs scan CT of the nasopharynx.
HISTORY: Head and neck cancer.
TECHNIQUE: Contiguous 2.5 mm axial images of the nasopharynx were performed without IV contrast. COMPARISON: xx/xx/xxxx.
FINDINGS: No soft tissue masses are seen within the soft tissues of the neck. The parotid and submandibular glands are predominantly fatty-replaced. Soft tissues of the Naso, oropharynx are unremarkable. There may be mild fat stranding of the right parapharyngeal soft tissues (series 1001, image 32). No abnormal masses are seen at that site. No bulky lymphadenopathy is seen. There is a fusiform aneurysm of the basilar artery as previously described. It appears to the mildly increased in size and currently measures 2.0 cm in transverse dimensions and previously measured 1.8 cm. It measured 1.5 cm in transverse dimensions on xx/xx/xxxx. Atherosclerotic calcifications are also seen within the carotid arteries bilaterally. There is near-complete opacification of the maxillary sinuses bilaterally. This has increased predominantly within the left maxillary sinus and mildly within the right maxillary sinus. The ethmoidal air cells are clear. Sphenoidal and frontal sinuses are clear. Degenerative changes of the cervical spine are noted.
IMPRESSIONS: 1. No soft tissue masses however, mild right parapharyngeal fat stranding is seen it may be postoperative or post radiation in nature. 2. Basilar artery aneurysm that has gradually increased in size when compared to prior examinations. 3. Atherosclerotic disease of the coronary arteries bilaterally..

Figure 1, Shin, H. et al. (2016)

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

DEEPHEALTH

De-identification methodology

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Target Detection

Methods

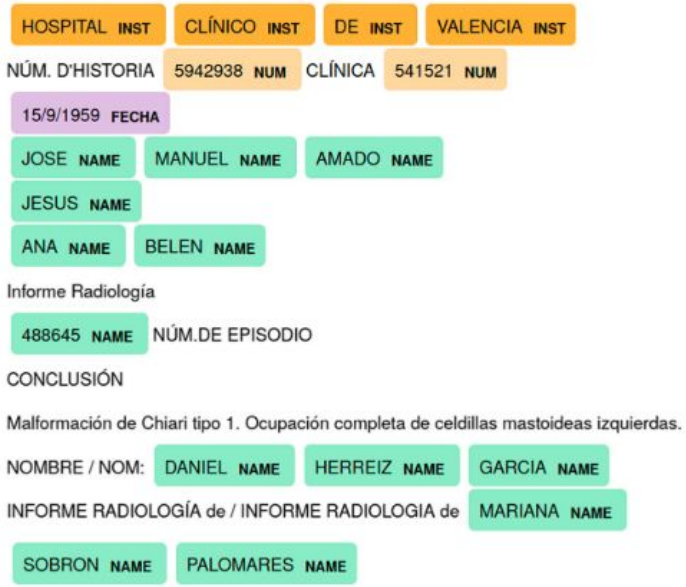


Figure 3b, Pérez-Díez, I. et al. (2021)

- Manual Detection
- Automatic Detection
  - Regex, lists of terms,...
    - dd/mm/yyyy, dd.mm.yyyy
    - mm/dd/yyyy, mm.dd.yyyy
  - Machine Learning Named Entity Recognition
    - Name, Date, Number,...

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Target Detection

Regular expression and lists of terms matching

DEEPHEALTH

Matching dates

^([1-9] |1[0-9]| 2[0-9]|3[0-1])(.|-)([1-9] |1[0-2])(.|-|)20[0-9][0-9]$

More Specific pattern

^(?:(?:31(\/|-|\.)(?:0?[13578]|1[02]|(?:Jan|Mar|May|Jul|Aug|Oct|Dec
)))\1|(?:(?:29|30)(\/|-|\.)(?:0?[1,3-9]|1[0-2]|(?:Jan|Mar|Apr|May|Jun|J
ul|Aug|Sep|Oct|Nov|Dec))\2))(?:(?:1[6-9]|[2-9]\d)?\d{2})$|^(?:29(\/|-
|\.)(?:0?2|(?:Feb))\3(?:(?:(?:1[6-9]|[2-9]\d)?(?:0[48]|[2468][048]|[13
579][26])|(?:(?:16|[2468][048]|[3579][26])00))))$|^(?:0?[1-9]|1\d|2[0
-8])(\/|-|\.)(?:(?:0?[1-9]|(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep))|(
?:1[0-2]|(?:Oct|Nov|Dec)))\4(?:(?:1[6-9]|[2-9]\d)?\d{2})$

List of existing names

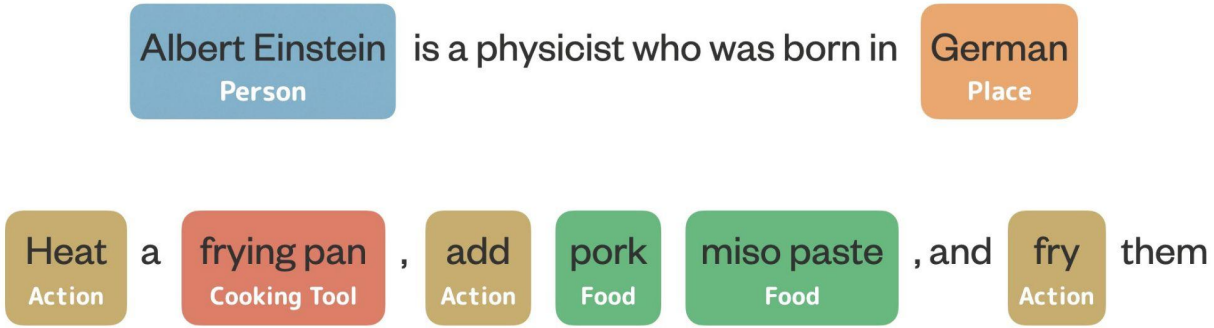
Nombre	Frecuencia
MARIA CARMEN	647,877
MARIA	589,055
CARMEN	375,835
ANA MARIA	271,616
JOSEFA	262,571
MARIA PILAR	260,302
ISABEL	258,873
LAURA	257,680
MARIA DOLORES	255,040
MARIA TERESA	248,131
ANA	246,892
CRISTINA	228,022
MARTA	226,974
MARIA ANGELES	223,560
LUCIA	208,878
FRANCISCA	204,718
MARIA ISABEL	203,839
MARIA JOSE	203,236

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Target Detection

Named Entity Recognition (NER)

NER Definition - *The task of identifying and categorizing key information (entities) in text. NER is a form of Natural Language Processing (NLP).*



https://medium.com/optuna/nerman-named-entity-recognition-system-built-on-allennlp-and-optuna-c044c319b955

Examples of NERs

- Names
- Organizations
- Places
- Phone Numbers
- Genes
- Diseases
- Compounds

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

De-identification

MRI Brain
Date of service: 24/11/2020
Patient: Alberto Pérez
Findings: The ventricles, cisterns...

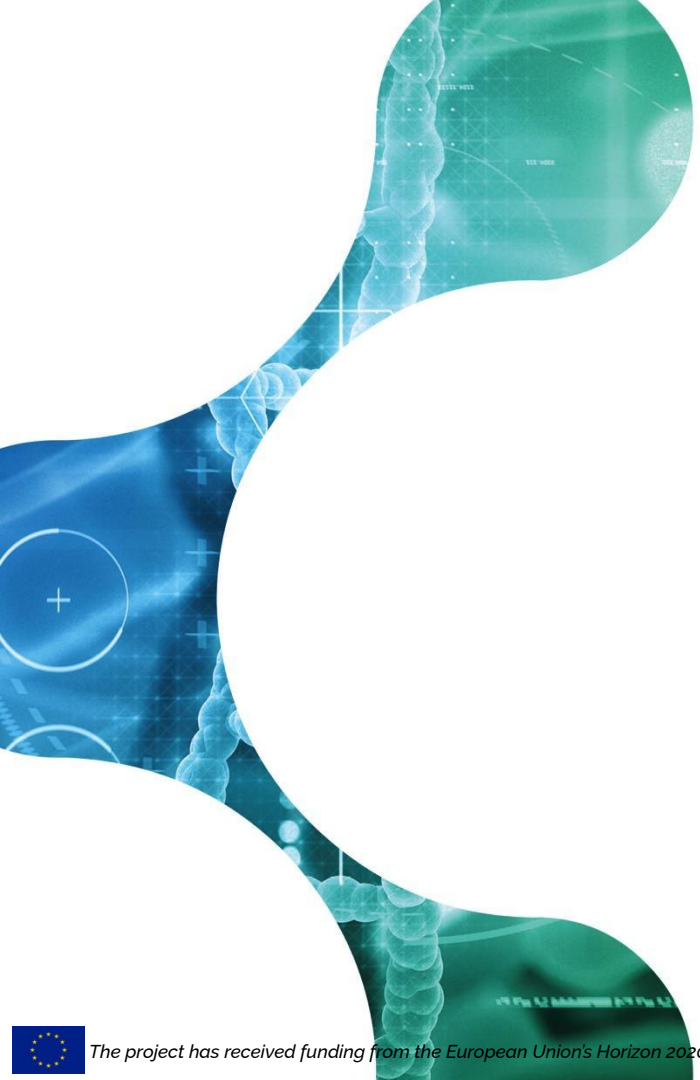
Substitution

MRI Brain
Date of service: 13/04/2002
Patient: Juan Martínez
Findings: The ventricles, cisterns...

Elimination

MRI Brain
Date of service: XX/XX/XXXX
Patient: XXXXX
Findings: The ventricles, cisterns...

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.



DEEPHEALTH

Named Entity Recognition

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

NER - Machine Learning

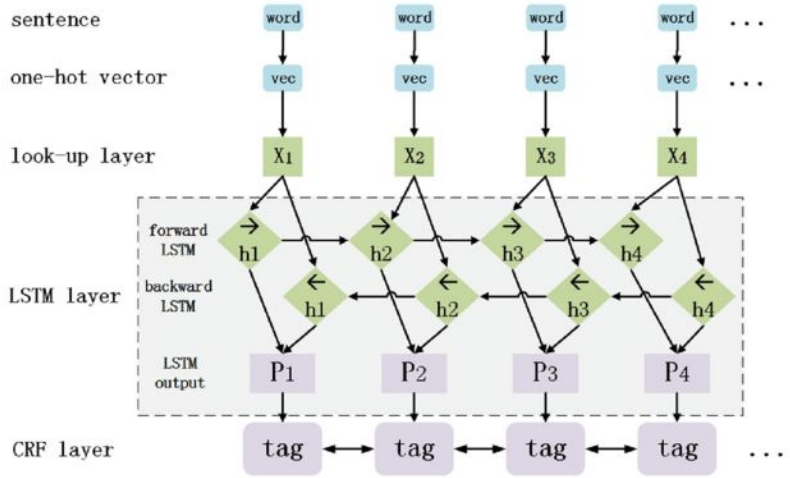
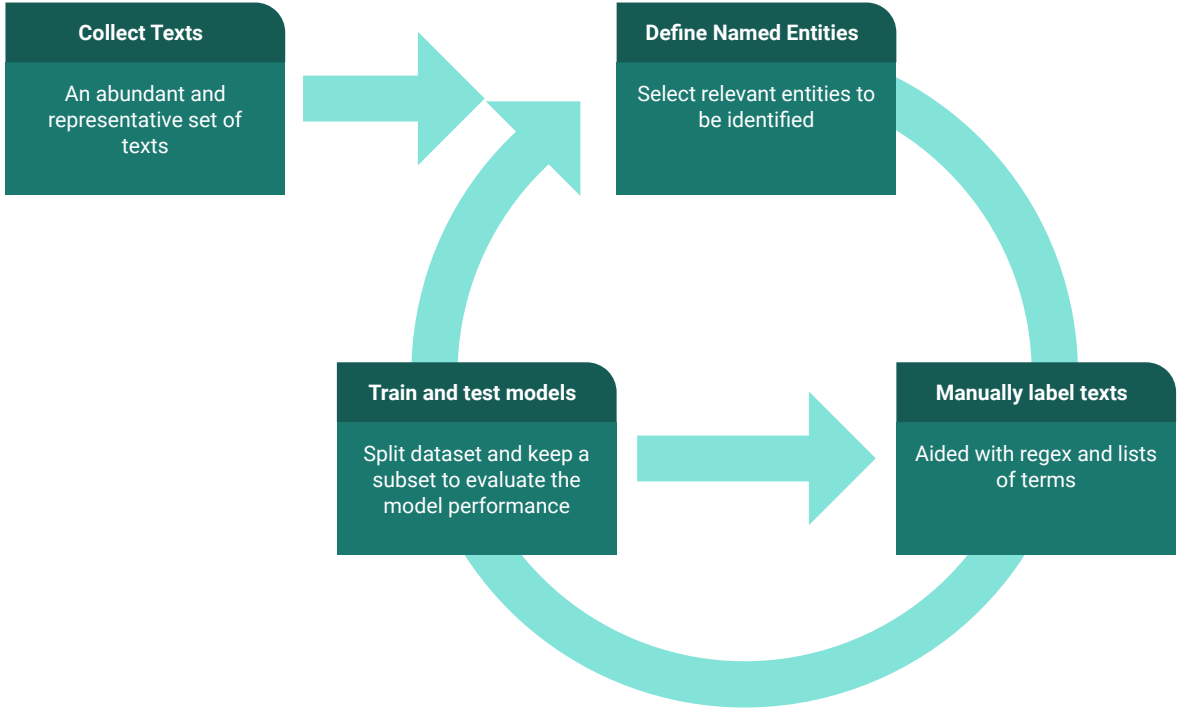


Figure 3, Ji, B. et al. (2019)

- Automatic processing of texts
- Fast and Accurate
- Adaptable to custom entities
- Requires a ML model trained with a suitable annotated dataset (corpus)
- Generation of corpus is expensive in terms of time and work

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Corpus construction



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

NER metrics

	Predicted Positives	Predicted Negatives
Positives	True Positives (TP)	False Negatives (FN)
Negatives	False Positives (FP)	True Negatives (TN)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Applications of NER

Texts Anonymization

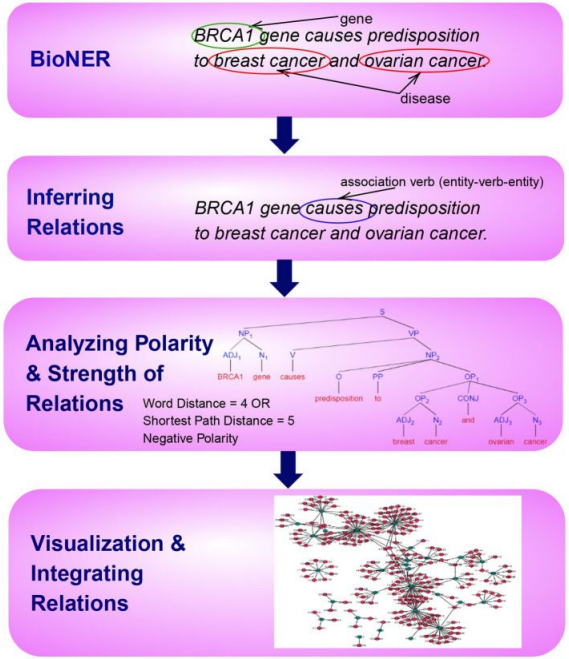
MRI Brain
Date of service: XX/XX/XXXX
Patient: XXXXX
Findings: The ventricles, cisterns...

Structure Data

MRI Brain
Date of service: 24/11/2020
Patient: Alberto Pérez
Findings: The ventricles, cisterns...

Patient	Alberto Pérez
Date	24/11/2020
Disease	...
Treatment	...

Information Extraction



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

DEEPHEALTH

Our de-identification method

Dismed

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

The dataset

A

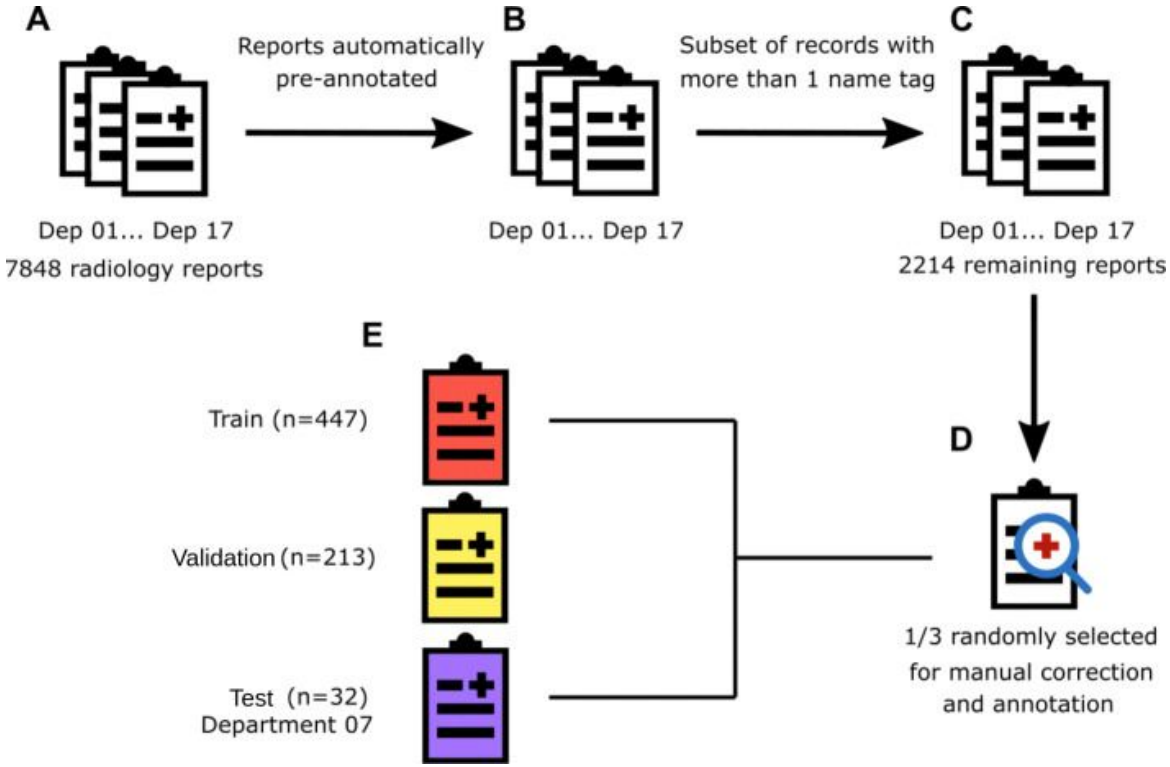
DIAGNÒSTIC / DIAGNÓSTICO
TORRES NAME
NOM FACULTATIU / NOMBRE FACULTATIVO
RAFAEL NAME TORRES NAME NAVAJAS NAME
METGE PETICIONARI / MÉDICO PETICIONARIO
JOSE NAME ALONSO NAME GIL NAME
R.M. Cerebral, 26/07/2017 FECHA
R.M. Columna cervical, 08/04/2011 FECHA
VALORACIÓ CAB / VALORACIÓN CAB
protocolo de volumetría cerebral y cervical, gadovist 10 ml iv.
ggc-38139
Tras valoración comparativa con previos, ausencia de incremento de carga lesional o de actividad inflamatoria, no existiendo focos con realce tras el contraste.
Volumen C2-C6: 6.788cm3
área C2. 94.67m2
Hospital INST Clínico INST Universitario INST de INST Granada INST
SIP 294321 NUM NÚM.D'HISTÒRIA CLÍNICA
NÚM. DE HISTORIA CLÍNICA 594637 NUM
DATA NAIXEMENT
FECHA NACIMIENTO 13 de enero de 1972 FECHA
DIRECCIÓ
DIRECCION CALLE DIR PAJARO DIR VERDE DIR - 21 18299 DIR
VALENCIA LOC

B

Datos del paciente
Nombre: Rocío NAME
Apellidos: Pérez NAME Ontiveros NAME
NHC: 22 75689632 36 NUM
Domicilio: Av DIR de DIR Leon DIR 66 DIR 1H DIR
Localidad/Provincia: Lleida LOC
CP: 06233 NUM
Fecha de nacimiento: 05/04/1937 FECHA
País de nacimiento: España LOC
Edad: 10 años Sexo: Varón
Fecha de ingreso: 15/08/2016 FECHA
Servicio: Oftalmología
Médico: Ender NAME Goñi NAME Moreno NAME NºCol: 15 15 31525 NUM
Consulta por dolor abdominal, observándose en la ecografía una masa renal. Se realizó biopsia tru-cut diagnosticada de tumor mesenquimal benigno. Fue intervenido quirúrgicamente.
Hallazgos CAB histológicos CAB
Se trataba de una proliferación de células fusocelulares con zonas de diferentes densidades sin atípias ni mitosis. Las células del estroma eran positivas para CD-34 y vimentina.
Ultraestructuralmente el tumor presenta células mesenquimales inmaduras
Dirección para correspondencia: Irene NAME Amat NAME Villegas NAME Pedro DIR de DIR Alejandría DIR Nº 1 DIR , 31014 DIR Pamplona LOC Navarra LOC

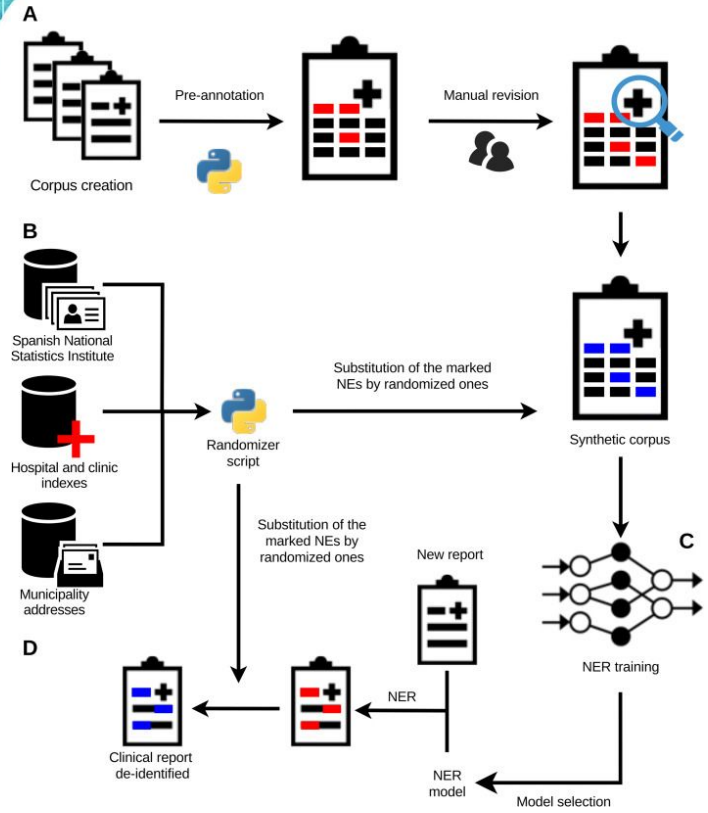
The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Corpus Construction



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Dismed pipeline



NEs	Description
CAB	Section headers
NAME	Names and surnames (patient and others)
DIR	Full addresses, including streets, numbers and zip codes
LOC	Cities, inside and outside addresses
NUM	Numbers or alphanumeric strings that might identify someone, including digital signatures, patient numbers, medical numbers, medical license numbers and others
FECHA	Dates
INST	Hospitals, healthcare centres or other institutions that might point to someone's location

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Results

	Training set			Validation set			Test set			MEDDOCAN		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
CAB	99.43	96.54	97.96	98.28	93.98	96.08	92.54	74.49	82.52	4.76	33.33	8.33
DIR	100	100	100	94.28	63.96	76.01	87.79	74.77	61.46	43.15	4.47	8.01
FECHA	100	100	100	98.54	99.04	98.78	98.20	97.53	97.86	51.39	89.41	65.13
INST	99.97	99.96	99.98	98.19	97.24	97.71	93.50	98.00	95.69	45.72	12.28	19.27
LOC	100	100	100	76.64	54.66	63.80	61.04	26.85	36.79	7.19	0.32	0.59
NAME	100	99.99	99.99	98.34	98.28	98.31	88.78	94.29	93.19	75.62	83.91	79.23
NUM	100	100	100	97.81	95.65	96.72	95.11	87.56	91.18	68.50	60.32	63.99
	99.87	99.28	99.58	98.06	96.10	97.08	93.23	89.39	91.31	65.63	55.37	59.98

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Publication

Pé *et al. Journal of Biomedical Semantics* (2021) 12:6
https://doi.org/10.1186/s13326-021-00236-2

Journal of Biomedical Semantics

RESEARCH Open Access

De-identifying Spanish medical texts - named entity recognition applied to radiology reports

Irene Pérez-Díez[1,2†], Raúl Pérez-Moraga[1,3†], Adolfo López-Cerdán[1,2], Jose-Maria Salinas-Serrano[4] and María de la Iglesia-Vayá[1,5,6*]

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

DEEPHEALTH

Thank you for your attention!

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.