ARPHAI
Gestión epidemiológica basada en inteligencia artificial y ciencia de datos

CIECTI
Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación

Global South AI4COVID Program

IDRC · CRDI
International Development Research Centre
Centre de recherches pour le développement international
Canada

Sida
SWEDISH INTERNATIONAL DEVELOPMENT COOPERATION AGENCY

Desarrollo y análisis de sesgos de un modelo de desidentificación de historias clínicas electrónicas en español

Sabrina Laura López*, Mariela Rajngewerc*, Luciano Silvi, Laura Ación, Laura Alonso Alemany

ic
instituto de cálculo
UBA - CONICET

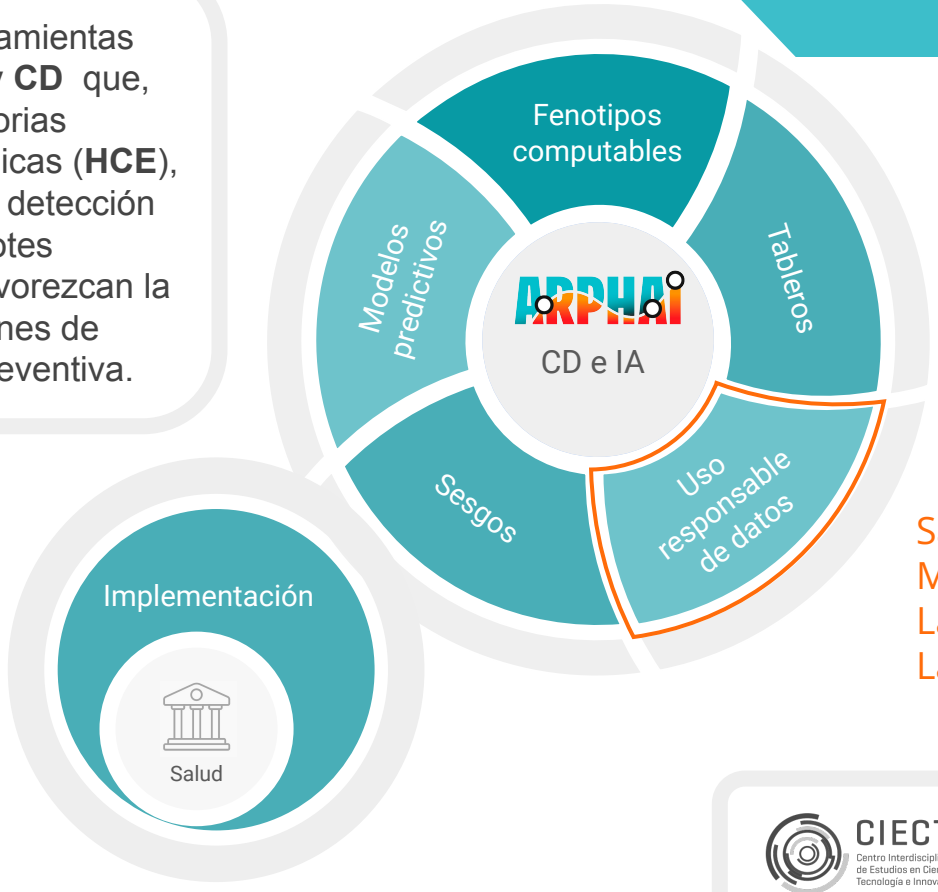
FAMAF
Facultad de Matemática, Astronomía, Física y Computación

UNC
Universidad Nacional de Córdoba

CONICET

ARPHAI

Desarrollar herramientas basadas en **IA y CD** que, aplicadas a historias clínicas electrónicas (**HCE**), contribuyan a la detección temprana de brotes epidémicos y favorezcan la toma de decisiones de salud pública preventiva.



Sabrina López
Mariela Rajngewerc
Laura Ación
Laura Alonso Alemany

CIECTI
Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación

Salud

Ciencia y Tecnología

Introducción

Uso secundario

Datos sensibles

Ley **Nº 25.326** de **Protección de los Datos Personales**

Protegidos nacional e internacionalmente **SENSIBLES**

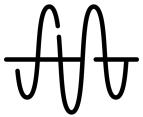
Impacto económico y/o moral en la vida de las personas

↓

Por ej. Discriminación en el mercado laboral

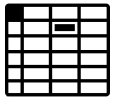
Datos de salud

Imágenes



Señales

Datos genómicos



Atributos

Datos geoespaciales

Texto Libre

Grafos

Ejemplos de evoluciones

Nombre: Cecilia Grierson Edad: 24 años Tel: 2745340887 OS: No DNI: 12345678
PT:75 kg peso Talla: 1.63 cm Antecedentes: familia materna diabetes…

Visita a domicilio Av. Rivadavia 742. Asisten dra. Ayelén Perez MP1234 y agente sanitario Nehuen Vilca

Ejemplos de evoluciones

Nombre: Cecilia Grierson Edad: 24 años Tel: 2745340887 OS: No DNI: 12345678
PT:75 kg peso Talla: 1.63 cm Antecedentes: familia materna diabetes…

Visita a domicilio Av. Rivadavia 742. Asisten dra. Ayelén Perez MP1234 y agente sanitario Nehuen Vilca

paciente con uso problem de sustancias. PN 01/01/21

osteooprosis y fragilidad osea sosp sme de Bruck

Prevalencia: 6-12%

Ejemplos de evoluciones

Nombre: Cecilia Grierson Edad: 24 años Tel: 2745340887 OS: No DNI: 12345678
PT:75 kg peso Talla: 1.63 cm Antecedentes: familia materna diabetes…

Visita a domicilio Av. Rivadavia 742. Asisten dra. Ayelén Perez MP1234 y agente sanitario Nehuen Vilca

paciente con uso problem de sustancias. PN 01/01/21

osteooprosis y fragilidad osea sosp sme de Bruck

Prevalencia: 6-12%

¡Imposibilidad de anotación manual!

El problema

Es un problema de **PLN** de **reconocimiento de entidades** de **Información Personal Protegida (IPP)**.

Visita a domicilio **Av. Rivadavia 742**. Asisten dra. **Ayelen Perez** MP**1234** y agente sanitario **Nehuen Vilca**

Visita a domicilio **XXX**. Asisten dra. **XXX** MP**XXX** y agente sanitario **XXX**

El problema

Es un problema de **PLN** de **reconocimiento de entidades** de **Información Personal Protegida (IPP)**.

En un **dominio específico** y muy particular: registros de atención de la salud.

- Frases no gramaticales.
- Errores de tipeo.
- Frecuente uso de siglas (cns, aco, csv, etc.)
- Ambigüedades.

El problema

Es un problema de **PLN** de **reconocimiento de entidades** de **Información Personal Protegida (IPP)**.

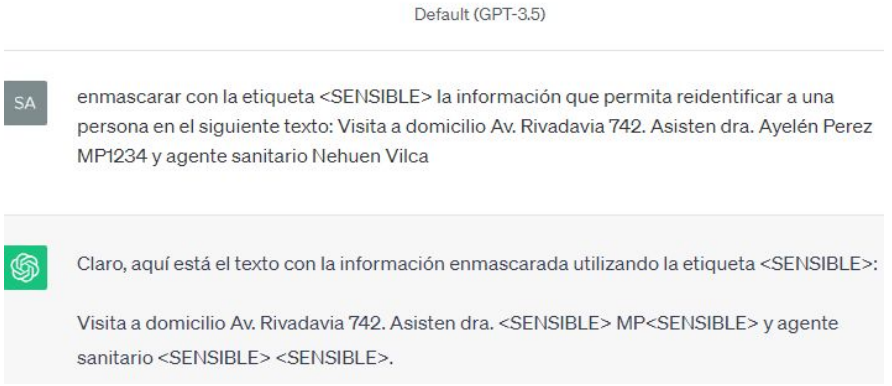
En un **dominio específico** y muy particular: registros de atención de la salud.

- Frases no gramaticales.
- Errores de tipeo.
- Frecuente uso de siglas (cns, aco, csv, etc.)
- Ambigüedades.

Modelos entrenados mayoritariamente en inglés. En español, sólo casos clínicos enriquecidos en IPP o informes de imágenes.

Modelos

https://huggingface.co/PlanTL-GOB-ES/es_anonimization_core_lg



https://chat.openai.com/

Modelos

```
import scrubadub, scrubadub_spacy
import spacy_transformers
scrubber = scrubadub.Scrubber(locale='es_AR')
scrubber.add_detector(scrubadub_spacy.detectors.SpacyEntityDetector(model ='es_core_news_md'))
```

```
text = 'Visita a domicilio Av. Rivadavia 742. Asisten dra. Ayelen Perez MP1234 y agente sanitario Nehuen Vilca'
```

```
scrubber.clean(text)
```

```
'Visita a domicilio Av. Rivadavia 742. Asisten dra. {{NAME}} y agente sanitario {{NAME}}'
```

```
scrubber = scrubadub.Scrubber(locale='es_AR')
scrubber.add_detector(scrubadub.detectors.TextBlobNameDetector)
scrubber.clean(text)
```

```
'{{NAME}} a domicilio Av. {{NAME}} 742. {{NAME}} dra. {{NAME}} {{NAME}} MP1234 {{NAME}} agente sanitario {{NAME}} {{NAME}}'
```

https://scrubadub.readthedocs.io/en/stable/

Modelos

```
import spacy
nlp = spacy.load("es_core_news_md")
text = "Visita a domicilio Av. Rivadavia 742. Asisten dra. Ayelen Perez MP1234 y agente sanitario Nehuen Vilca"

# Process the text
doc = nlp(text)

# Iterate over the entities
for ent in doc.ents:
    # Print the entity text and label
    print(ent.text, ent.label_)
```

```
Av. Rivadavia LOC
Asisten dra MISC
Ayelen Perez MP1234 PER
Nehuen Vilca PER
```

spaCy

https://spacy.io/

¿Qué tipo de errores pueden cometer los modelos?

Input: Visita a domicilio **Av. Rivadavia 742**. Asisten dra. **Ayelen Perez** MP**1234** y agente sanitario **Nehuen Vilca**

Output: **XXX** a domicilio **XXX**. Asisten dra. **XXX** MP**XXX** y agente sanitario **XXX**

❌ ✅ ✅ ✅ ✅ ✅ ✅ ✅ ✅ ✅ ✅ ✅

⇨ El modelo considera que la palabra “Visita” es una entidad identificatoria por lo que elimina una porción del texto que no era un dato identificatorio. Llamaremos a este tipo de error **Falso Positivo**.

¿Qué tipo de errores pueden cometer los modelos?

Input: Visita a domicilio **Av. Rivadavia 742**. Asisten dra. **Ayelen Perez** MP**1234** y agente sanitario **Nehuen Vilca**

Output: Visita a domicilio **Av. Rivadavia 742**. Asisten dr. **XXX** MP**XXX** y agente sanitario **XXX**

✅ ✅ ✅ ❌ ✅ ✅ ✅ ✅ ✅ ✅ ✅ ✅

⇨ El modelo considera que la expresión “**Av. Rivadavia 742**” no es una entidad identificatoria por lo que no elimina una porción del texto. Llamaremos a este tipo de error **Falso Negativo**.

Cómo evaluar los modelos de anonimización

Al cometer errores de tipo **Falso Negativo**, el modelo estará **exponiendo a sus usuarios** mientras que si comete errores de tipo Falso Positivo se estará eliminando información que no expone a las personas.

Para evaluar al modelo **consideraremos métricas que consideren y ponderen los Falsos Negativos**.

Métricas agregadas

- Evaluación general y permiten comparar modelos (*recall, accuracy, f1 score*)
- Limitaciones

Limitación de las métricas agregadas

Estas métricas pueden estar **ocultando la presencia de errores sistemáticos sobre subgrupos** de la población.

- Por ejemplo: si hay comunidades que cuentan con nombres poco frecuentes y los modelos fallan solamente sobre esos nombres, se estaría exponiendo a una población en particular y esto no podría ser detectado con las métricas agregadas.

Limitación de las métricas agregadas

Estas métricas pueden estar **ocultando la presencia de errores sistemáticos sobre subgrupos** de la población.

- Por ejemplo: si hay comunidades que cuentan con nombres poco frecuentes y los modelos fallan solamente sobre esos nombres, se estaría exponiendo a una población en particular y esto no podría ser detectado con las métricas agregadas.

Una forma de evaluar modelos que permite **detectar sesgos sistemáticos** sobre grupos de interés es mediante las **métricas de equidad**.

Recapitulando...

- Desarrollar una modelo automático basado en **expresiones regulares** y consulta de **diccionarios** para desidentificar texto libre de historias clínicas electrónicas.
  - Abierto.
  - Bajo costo computacional.
  - Interpretable.
  - Adaptable.

- Evaluar al algoritmo mediante **métricas agregadas** y **métricas de equidad**.

Metodología

Base de datos

- 2.394.499 registros de textos libre de
- 214.308 pacientes
- en el periodo 04/10/2016 - 28/01/2021
- Sistema público de atención ambulatoria

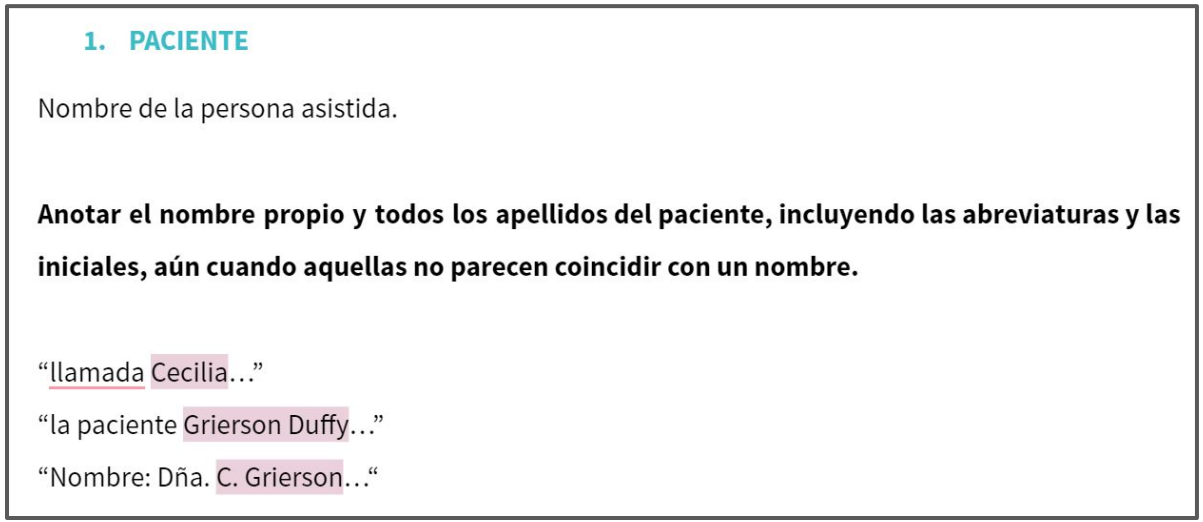
Anotación humana

- 22 categorías: PACIENTE, EDAD, GÉNERO, FAMILIAR, DRX, FECHA, EFECTOR, INSTITUCIÓN, DIRECCIÓN, ZONA, PAÍS, NÚM_TELÉFONO, CORREO_ELECTRÓNICO, NÚM_DNI, NÚM_CUIT_CUIL, PASAPORTE, MATRICULA, EPOF, PATENTE, NÚM_SERIE_DISPOSITIVOS, OTROS_NÚM y DUDOSOS.

Anotación humana

- 22 categorías: PACIENTE, EDAD, GÉNERO, FAMILIAR, DRX, FECHA, EFECTOR, INSTITUCIÓN, DIRECCIÓN, ZONA, PAÍS, NÚM_TELÉFONO, CORREO_ELECTRÓNICO, NÚM_DNI, NÚM_CUIT_CUIL, PASAPORTE, MATRICULA, EPOF, PATENTE, NÚM_SERIE_DISPOSITIVOS, OTROS_NÚM y DUDOSOS.

- Manual de anotación



1. **PACIENTE**

Nombre de la persona asistida.

Anotar el nombre propio y todos los apellidos del paciente, incluyendo las abreviaturas y las iniciales, aún cuando aquellas no parecen coincidir con un nombre.

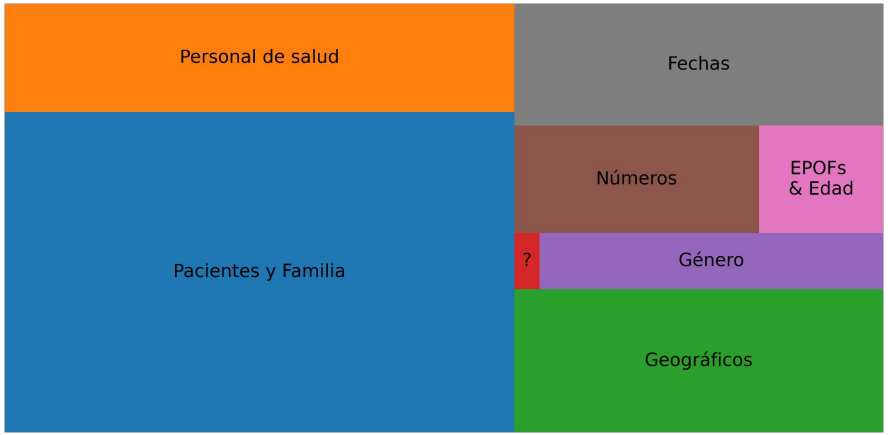
"llamada Cecilia…"
"la paciente Grierson Duffy…"
"Nombre: Dña. C. Grierson…"

Anotación humana

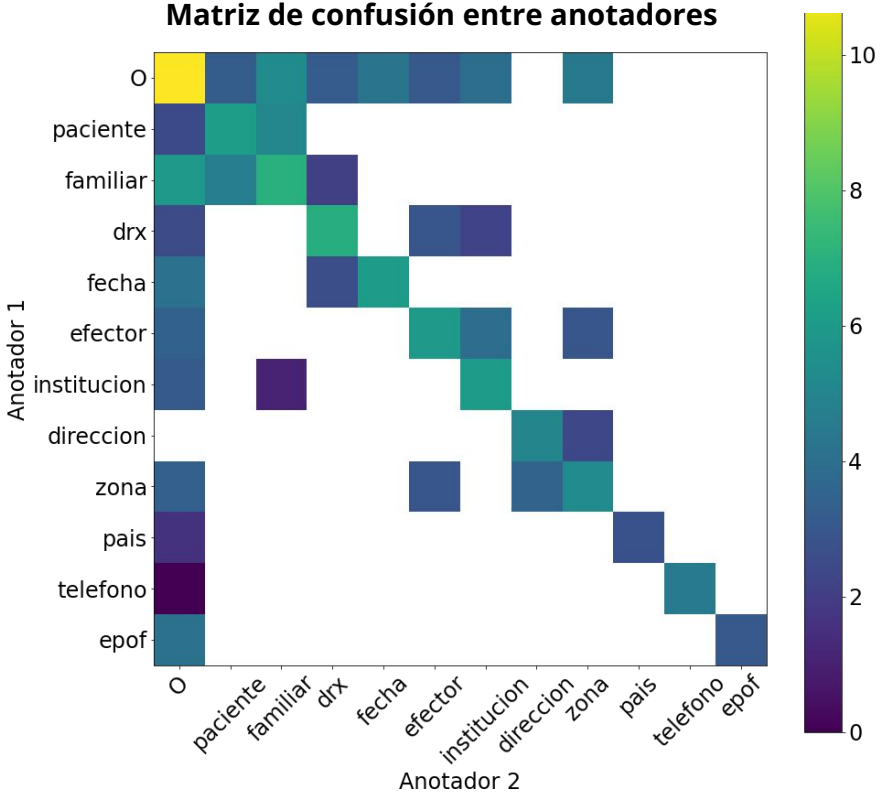
- 22 categorías: PACIENTE, EDAD, GÉNERO, FAMILIAR, DRX, FECHA, EFECTOR, INSTITUCIÓN, DIRECCIÓN, ZONA, PAÍS, NÚM_TELÉFONO, CORREO_ELECTRÓNICO, NÚM_DNI, NÚM_CUIT_CUIL, PASAPORTE, MATRICULA, EPOF, PATENTE, NÚM_SERIE_DISPOSITIVOS, OTROS_NÚM y DUDOSOS.
- Manual de anotación
- 2 anotadores
- 2500 registros a partir de muestreo aleatorio y dirigido
- 6 lotes de anotación con 10% de solapamiento

Resultados de anotación

Categorías de las 6111 entidades detectadas en 1442 registros



Matriz de confusión entre anotadores



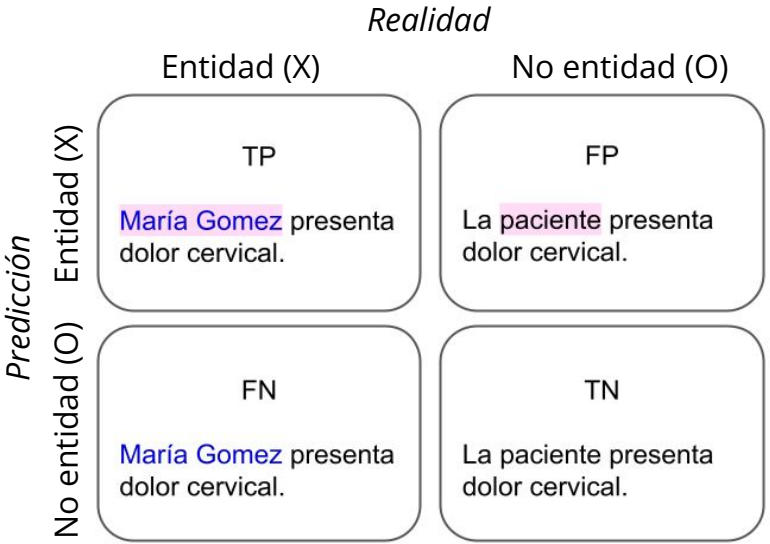
Modelo

- Expresiones regulares
  - correo electrónico = '[\w\.-]+@[\w\.-]+\.\w+'

- Diccionarios
  - Países, provincias, departamentos, localidades, municipios, calles.
  - Instituciones educativas
  - Hospitales y Centros de Salud
  - EPOFs
  - Nombres y apellidos

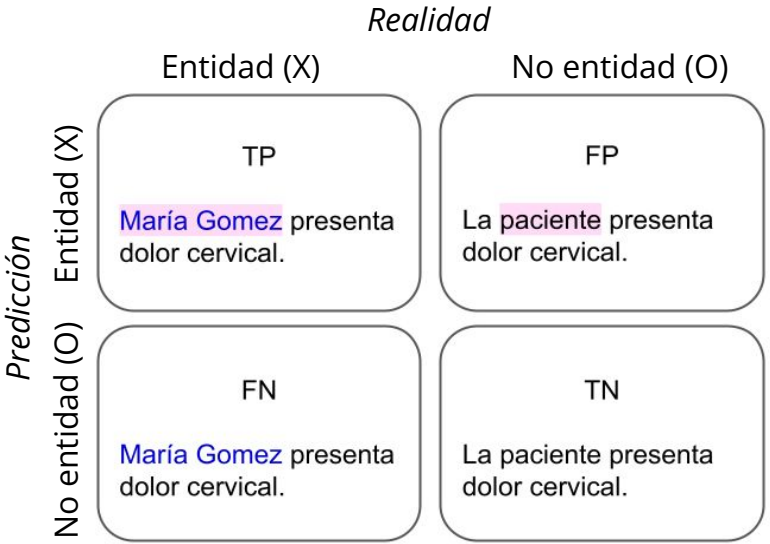
Dra. Perez → Dra. <PERSONAL DE SALUD>

Métricas agregadas



Se eligieron las métricas *recall, accuracy balanceado y f1 score* ya que contemplan los FN y permiten comparar el modelo

Métricas de equidad



- Treatment equality

$$\max_{i,j}\left\{\left|\frac{FN_i}{FP_i}-\frac{FN_j}{FP_j}\right|\right\}$$

- Equal Opportunity

$$\max_{i,j}\left\{\left|\frac{TP_i}{TP_i+FN_i}-\frac{TP_j}{TP_j+FN_j}\right|\right\}$$

Subpoblaciones de interés: **género** y **grupo etario**

Resultados

Evaluación del rendimiento del modelo de desidentificación

Sobre 1409 registros (9758 entidades y 78398 no-entidades).

Métrica de evaluación	Valor obtenido
Accuracy balanceado	0.76
Recall	0.56
F1 score	0.78

Visita a domicilio Av. Rivadavia 742 Asisten dra <DRX> MP <MATRICULA> y agente sanitario <DRX>

Análisis de sesgos por género

Matriz de confusión

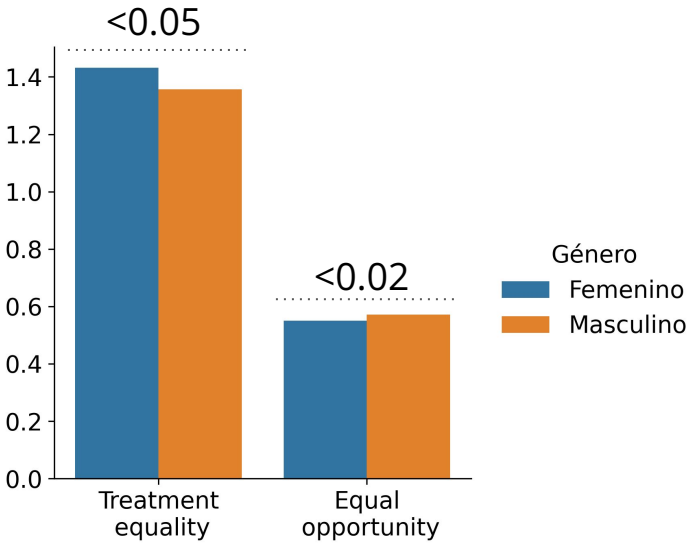
Predicción \ *Realidad*	Entidad (X)	No entidad (O)
Entidad (X)	TP María Gomez presenta dolor cervical.	FP La paciente presenta dolor cervical.
No entidad (O)	FN María Gomez presenta dolor cervical.	TN La paciente presenta dolor cervical.

- Treatment equality

$$\max_{i,j}\left\{\left|\frac{FN_i}{FP_i}-\frac{FN_j}{FP_j}\right|\right\}$$

- Equal Opportunity

$$\max_{i,j}\left\{\left|\frac{TP_i}{TP_i+FN_i}-\frac{TP_j}{TP_j+FN_j}\right|\right\}$$



Análisis de sesgos por edad*

Matriz de confusión

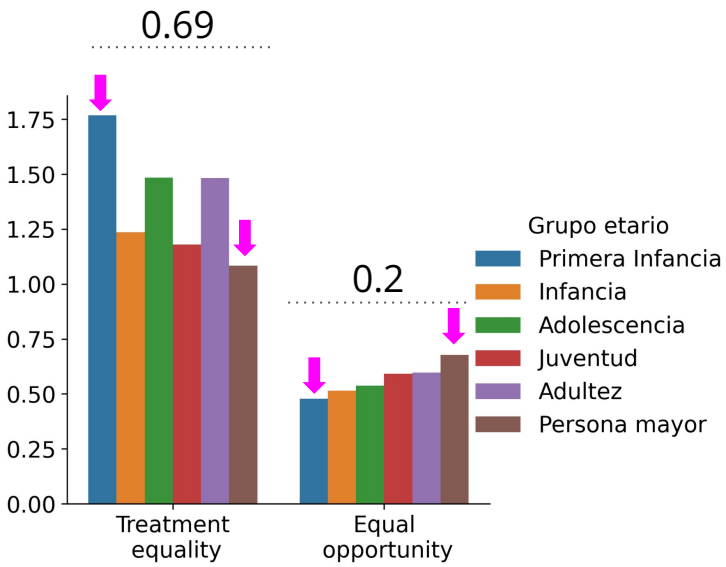
Predicción	*Realidad*	
	Entidad (X)	No entidad (O)
Entidad (X)	TP María Gomez presenta dolor cervical.	FP La paciente presenta dolor cervical.
No entidad (O)	FN María Gomez presenta dolor cervical.	TN La paciente presenta dolor cervical.

- Treatment equality

$$\max_{i,j}\left\{\left|\frac{FN_i}{FP_i}-\frac{FN_j}{FP_j}\right|\right\}$$

- Equal Opportunity

$$\max_{i,j}\left\{\left|\frac{TP_i}{TP_i+FN_i}-\frac{TP_j}{TP_j+FN_j}\right|\right\}$$



*Primera infancia (menor a 6 años), infancia (entre 6 y 11 años), adolescencia (entre 12 y 18 años), juventud (entre 19 y 26 años), adultez (entre 27 y 59 años) y personas mayores (mayores a 60 años).

Conclusiones

Conclusiones

- El **modelo propuesto** de bajo costo computacional, interpretable y adaptable, basado en expresiones regulares **permite identificar IPP**.
    - Resulta necesario continuar trabajando para mejorar la tasa de FN.

- **No** hemos observado **inequidades** respecto a los **géneros**.

- **Sí** hemos observado **inequidades** respecto a **segmentos etarios**.
    - pacientes de primera infancia sufren de una mayor exposición de IPP.

Trabajo futuro

Trabajo futuro

➔ Ampliar el análisis de inequidades
- problemas con la calidad del dato que define a las subpoblaciones

Trabajo futuro

- ➔ Ampliar el análisis de inequidades
  - problemas con la calidad del dato que define a las subpoblaciones
- ➔ Comparar el modelo propuesto con estrategias basadas en grandes modelos de lenguaje
  - actualmente estamos evaluando modelo basado en BETO.

Trabajo futuro

- ➔ Ampliar el análisis de inequidades
  - problemas con la calidad del dato que define a las subpoblaciones
- ➔ Comparar el modelo propuesto con estrategias basadas en grandes modelos de lenguaje
  - actualmente estamos evaluando modelo basado en BETO.
- ➔ Postprocesamiento
  - *hidden in plain sight*

Dra. Perez → Dra. Gomez

Trabajo futuro

- ➔ Ampliar el análisis de inequidades
  - problemas con la calidad del dato que define a las subpoblaciones
- ➔ Comparar el modelo propuesto con estrategias basadas en grandes modelos de lenguaje
  - actualmente estamos evaluando modelo basado en BETO.
- ➔ Postprocesamiento
  - *hidden in plain sight*

Dra. Perez → Dra. Gomez

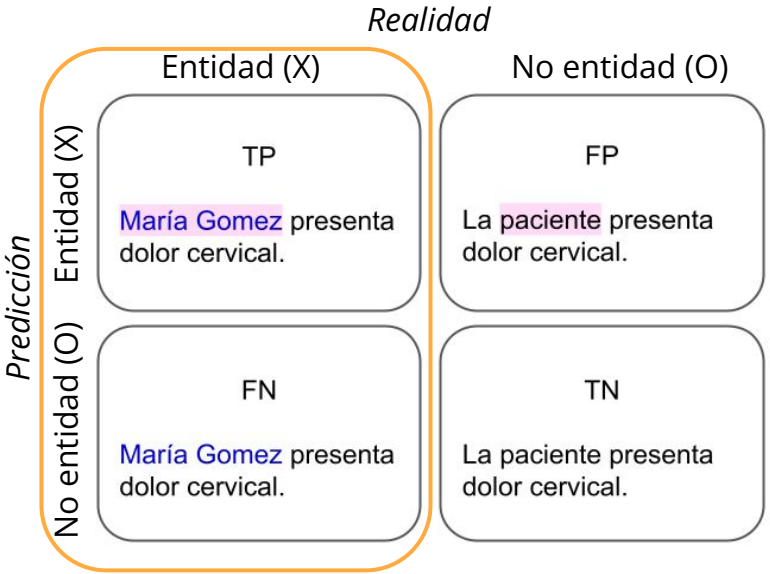
- ➔ Testearlo y mejorarlo sobre otros corpus (Anonimitaton)
  - parametrización de la localización

¡Muchas gracias!

Sabrina Laura López
Investigadora de la Línea Uso Responsable de Datos (ARPHAI)
Instituto de Cálculo, UBA-CONICET
www.ic.fcen.uba.ar
@SLLDeC sabrina.lopez.ds@gmail.com

Íconos de flaticon

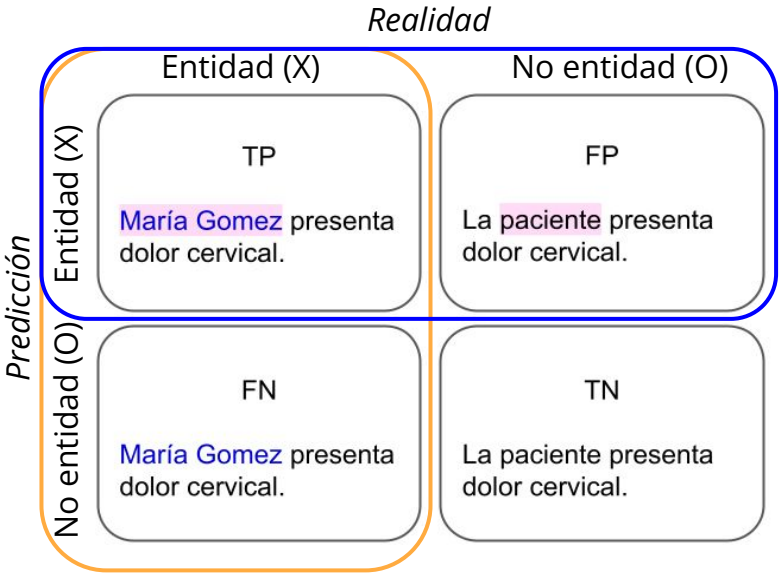
Métricas agregadas



entidades correctamente clasificadas respecto a los casos totales

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

Métricas agregadas



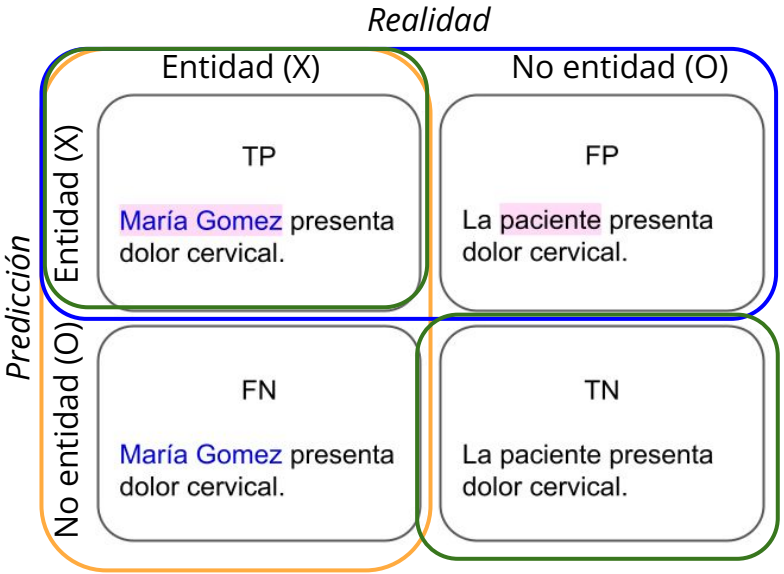
entidades correctamente clasificadas respecto a los casos totales

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

entidades correctamente clasificadas respecto al total de las predichas como entidades

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

Métricas agregadas



entidades correctamente clasificadas respecto a los casos totales

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

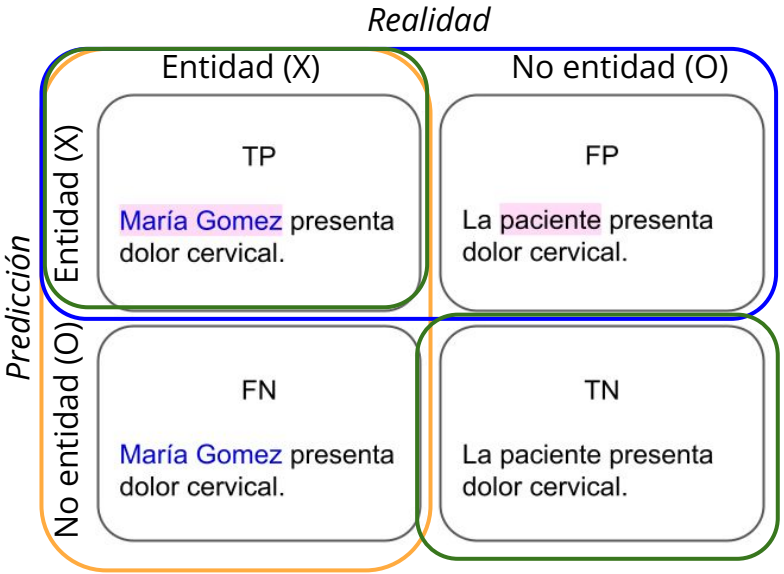
entidades correctamente clasificadas respecto al total de las predichas como entidades

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

casos correctamente clasificados respecto al total

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Métricas agregadas



entidades correctamente clasificadas respecto a los casos totales

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

entidades correctamente clasificadas respecto al total de las predichas como entidades

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

casos correctamente clasificados respecto al total

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

media armónica entre precision y recall

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