MDS Master of Data Science
Universidad de Chile

IDENTIFICACIÓN DE ENTIDADES EN PRESCRIPCIONES

CON EL OBJETIVO DE DETECTAR ERRORES DE MEDICACIÓN.

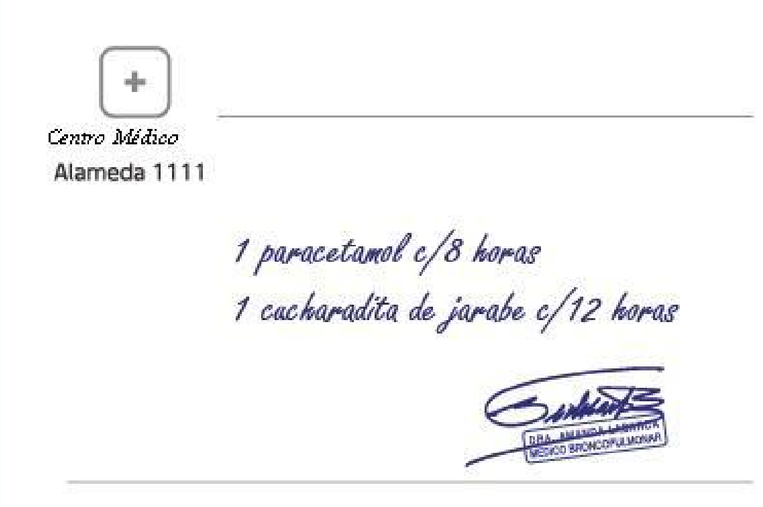
DANIEL CARMONA, MARTÍN SEPÚLVEDA,
MONSERRAT PRADO, CAMILO CARVAJAL

ENTIDADES MINSAL

TABLA DE CONTENIDO

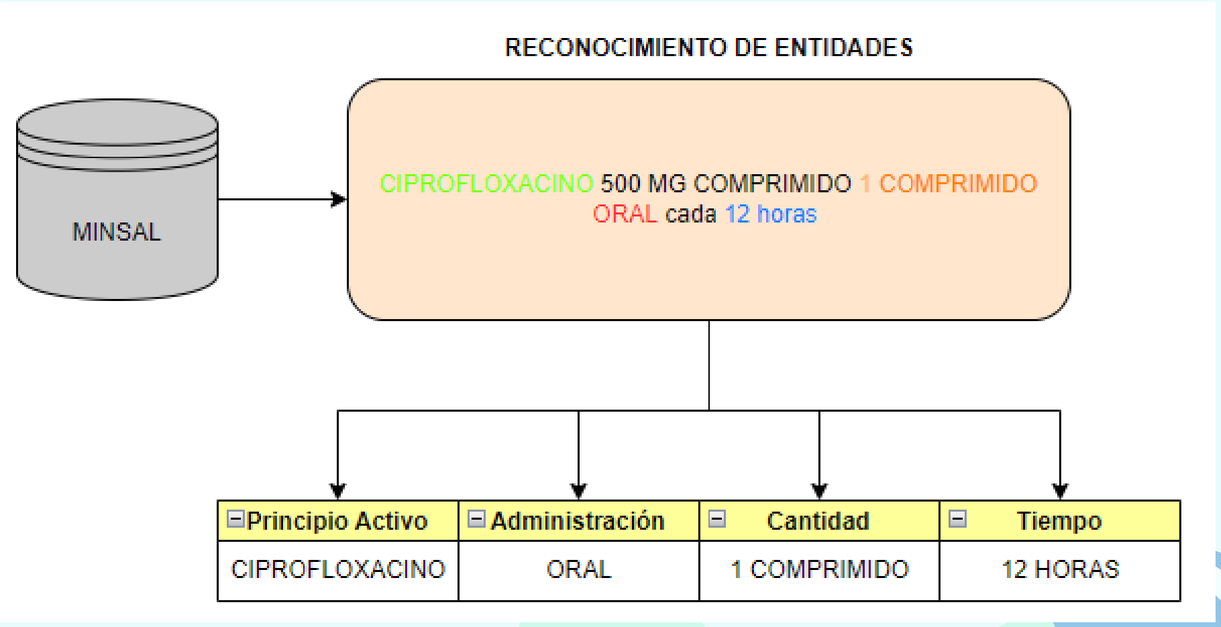
/1 DESCRIPCIÓN DEL PROBLEMA

- Existen recetas médicas que pueden carecer de cierta información importante, llevando a errores de medicación y a un empeoramiento en el estado del paciente.
- Las recetas electrónicas pueden contener campos de texto libre.
- Esto dificulta la verificación de la completitud de la prescripción.
- Reconocimiento de entidades facilita la detección de errores.



/1 DESCRIPCIÓN DEL PROBLEMA

- Dado un campo de texto libre, utilizar algoritmos de NLP para reconocer entidades.
- Detectar errores de completitud o gramática en las indicaciones.



/1

DESCRIPCIÓN DEL PROBLEMA

Un acercamiento a través de entidades en texto libre.

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL

PERIODICIDAD | DURACIÓN

cada 12 horas durante 15 dias

/1

DESCRIPCIÓN DEL PROBLEMA

Un acercamiento a través de entidades en texto libre.

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL

PERIODICIDAD | DURACIÓN

cada 12 horas durante 15 dias

/1

DESCRIPCIÓN DEL PROBLEMA

Un acercamiento a través de entidades en texto libre.

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

DOMPERIDONA 10 MG COMPRIMIDO 1 COMPRIMIDO

PERIODICIDAD

ORAL cada 6 horas

/1

DESCRIPCIÓN DEL PROBLEMA

Un acercamiento a través de entidades en texto libre.

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

DOMPERIDONA 10 MG COMPRIMIDO 1 COMPRIMIDO

PERIODICIDAD

ORAL cada 6 horas

/1

DESCRIPCIÓN DEL PROBLEMA

Un acercamiento a través de entidades en texto libre.

PRINCIPIO_ACTIVO

FORMA-FARMA

LEVETIRACETAM 100 MG/ML SOL. ORAL Suministro

ADMIN

PERIODICIDAD

inmediato primera vez. 1 FRASCO ORAL cada 12 horas

DURACIÓN

durante 12 horas.

/1

DESCRIPCIÓN DEL PROBLEMA

Un acercamiento a través de entidades en texto libre.

PRINCIPIO_ACTIVO

FORMA-FARMA

LEVETIRACETAM 100 MG/ML SOL. ORAL Suministro

ADMIN

PERIODICIDAD

inmediato primera vez. 1 FRASCO ORAL cada 12 horas

DURACIÓN

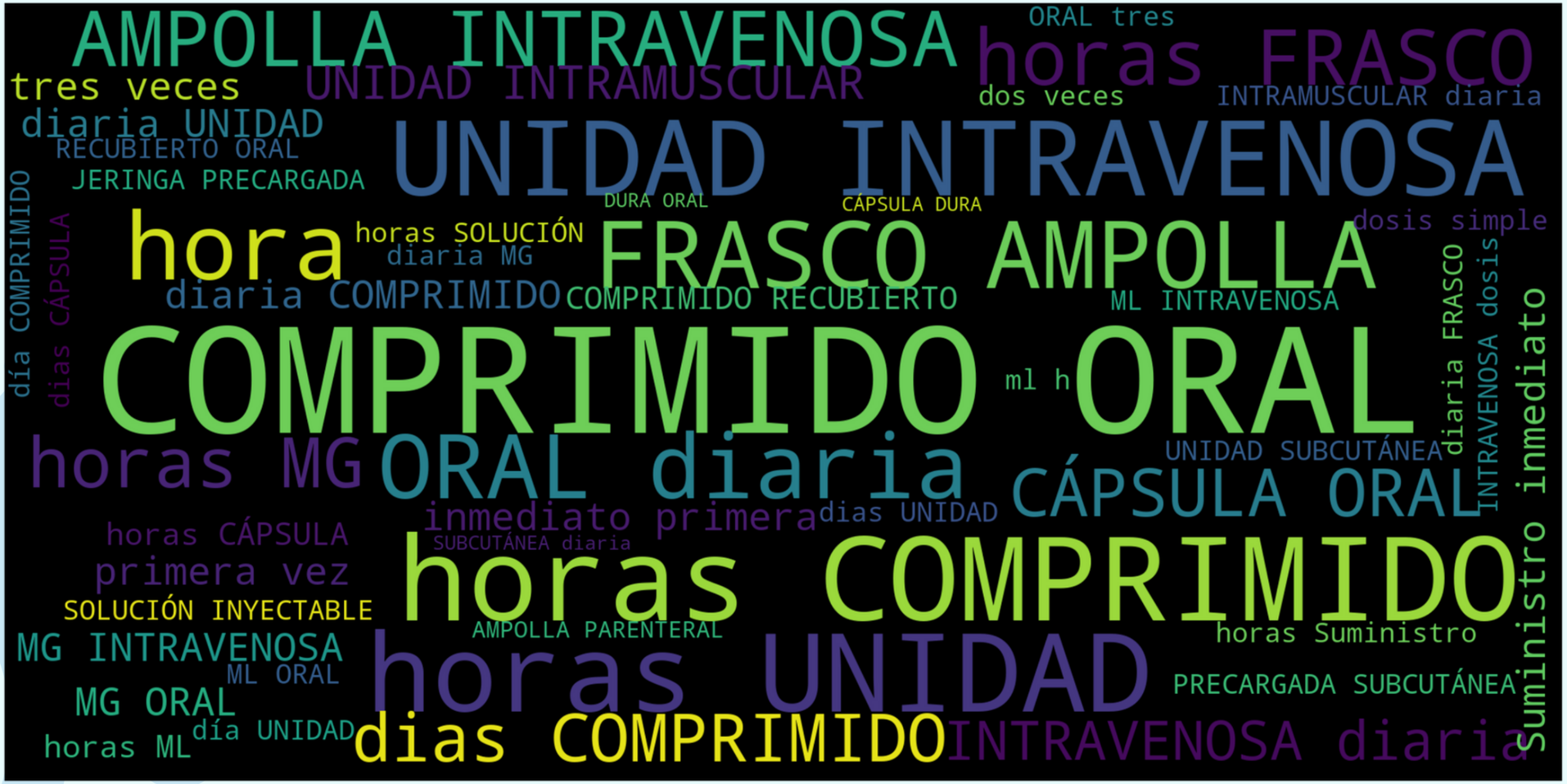
durante 12 horas.

/2 DATOS Y PREPROCESAMIENTO

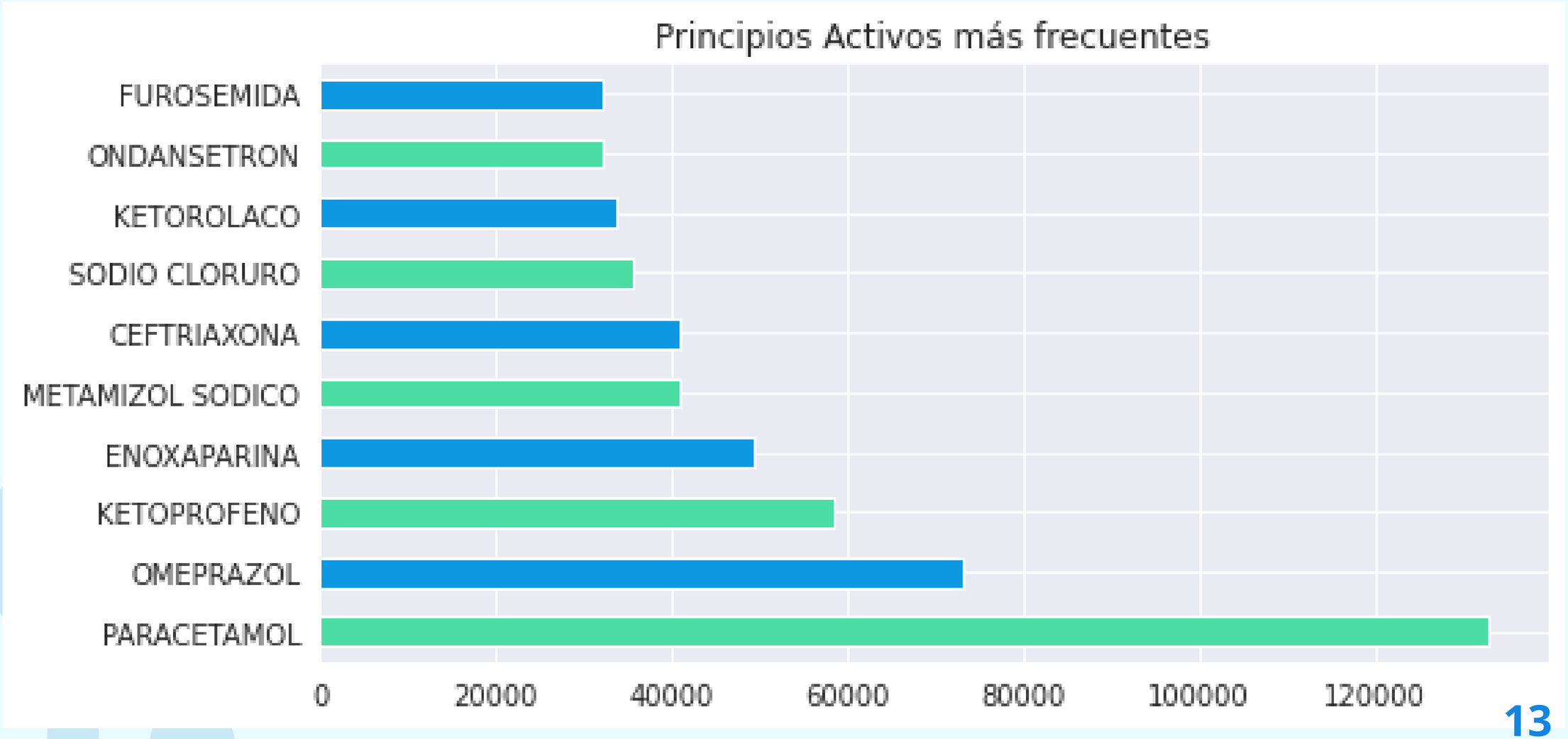
Un millón y medio de registros

/2 DATOS Y PREPROCESAMIENTO

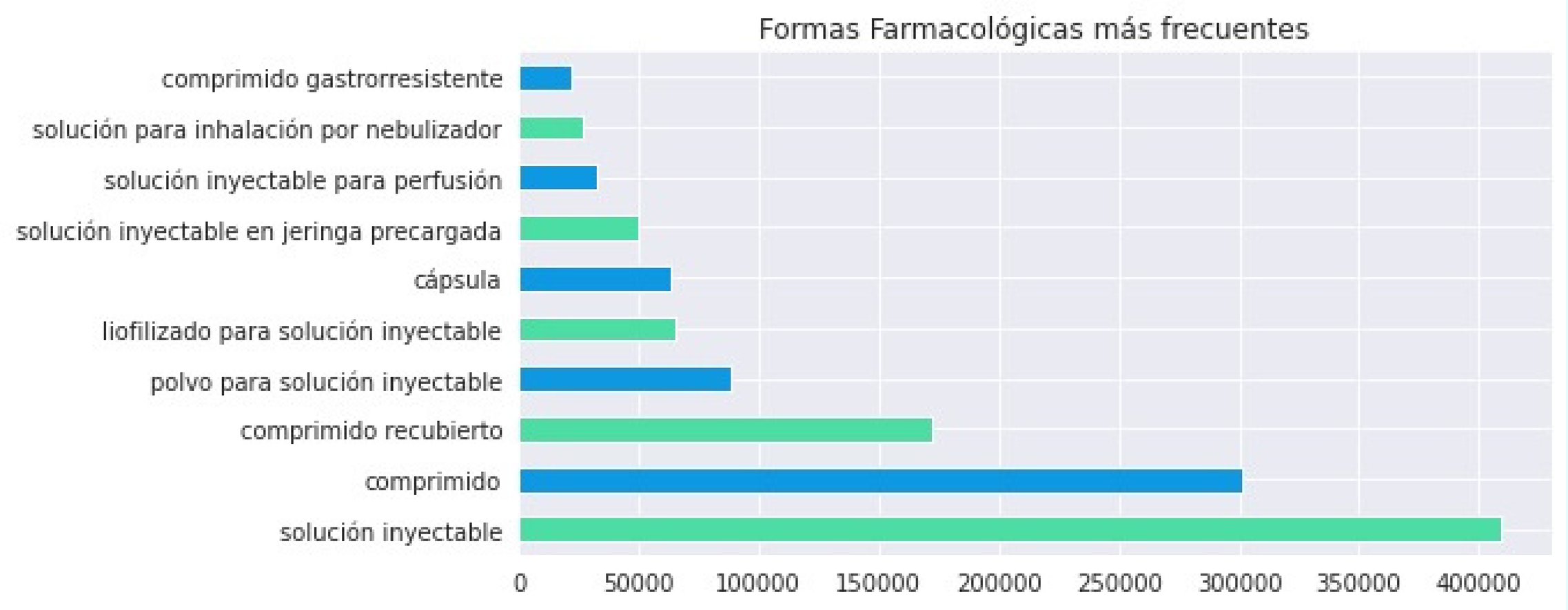
Resumen de la prescripción



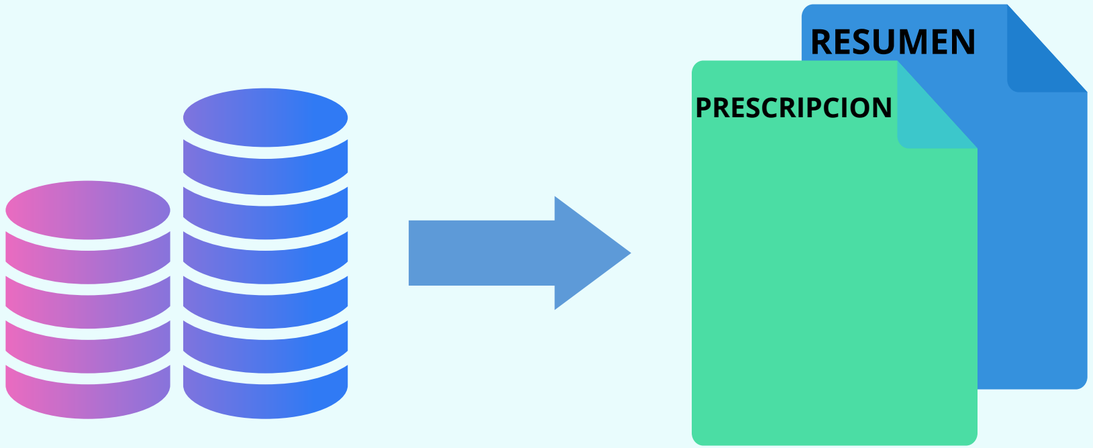
/2 DATOS Y PREPROCESAMIENTO



/2 DATOS Y PREPROCESAMIENTO



/2 DATOS Y PREPROCESAMIENTO



NÚMERO DE EJEMPLOS ÚNICOS SE REDUCE A 108.049

/2 DATOS Y PREPROCESAMIENTO

El problema se convierte en una clasificación para cada token en la secuencia.

PRINCIPIO_ACTIVO	FORMA-FARMA	ADMIN
HIDRALAZINA 50 MG	**COMPRIMIDO 13 MG**	**ORAL**

PERIODICITY	DURATION
cada 12 horas	**durante 15 dias**

/2 DATOS Y PREPROCESAMIENTO

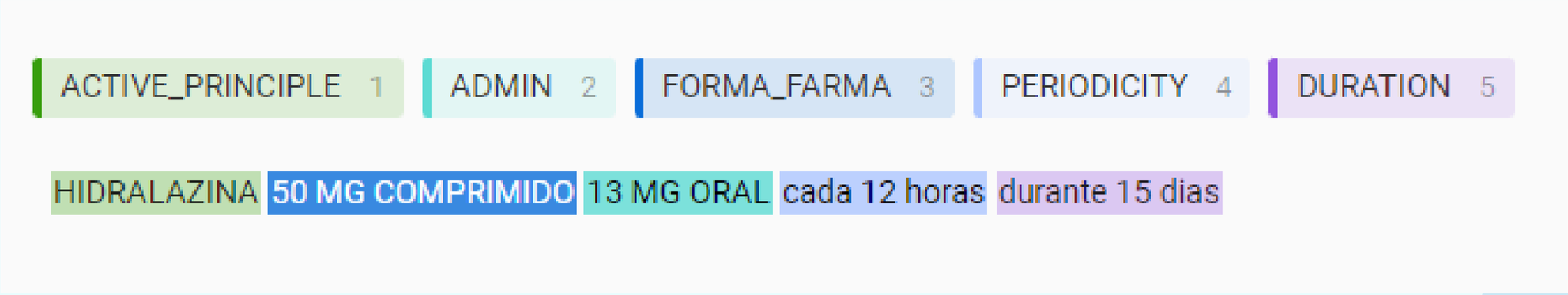
- Se etiquetan los datos a través de reglas.
- Se definen 5 entidades:
  - ACTIVE_PRINCIPLE
  - FORMA_FARMA
  - ADMIN
  - PERIODICITY
  - DURATION

PARACETAMOL	B-ACTIVE_PRINCIPLE
500	B-FORMA_FARMA
MG	I-FORMA_FARMA
COMPRIMIDO	I-FORMA_FARMA
1	B-ADMIN
COMPRIMIDO	I-ADMIN
ORAL	I-ADMIN
CADA	B-PERIODICITY
6	I-PERIODICITY
HORAS	I-PERIODICITY
DURANTE	B-DURATION
3	I-DURATION
DIAS	I-DURATION

/2 DATOS Y PREPROCESAMIENTO

ETIQUETADO DE DATOS MANUAL

- Se utilizó la herramienta Label Studio.
- Se etiquetaron 1000 recetas.



/3 MODELAMIENTO

MODELO REGEX

- Se utlizan 2 conjuntos de Principio Activo y Forma Farma
- Se reconocen expresiones regulares de Periodicidad , Duración y Admin.

['TRAMADOL'	'B-ACTIVE_PRINCIPLE']
['100'	'O']
['MG/ML'	'O']
['SOLUCIÓN'	'B-FORMA_FARMA']
['ORAL'	'I-FORMA_FARMA']
['FRASCO'	'O']
['10'	'O']
['ML'	'O']
['0,2'	'O']
['ML'	'O']
['ORAL'	'B-VIA_ADMIN']
['CADA'	'B-PERIODICITY']
['8'	'I-PERIODICITY']
['HORAS'	'I-PERIODICITY']
['DURANTE'	'B-DURATION']
['15'	'I-DURATION']
['DIAS'	'I-DURATION']

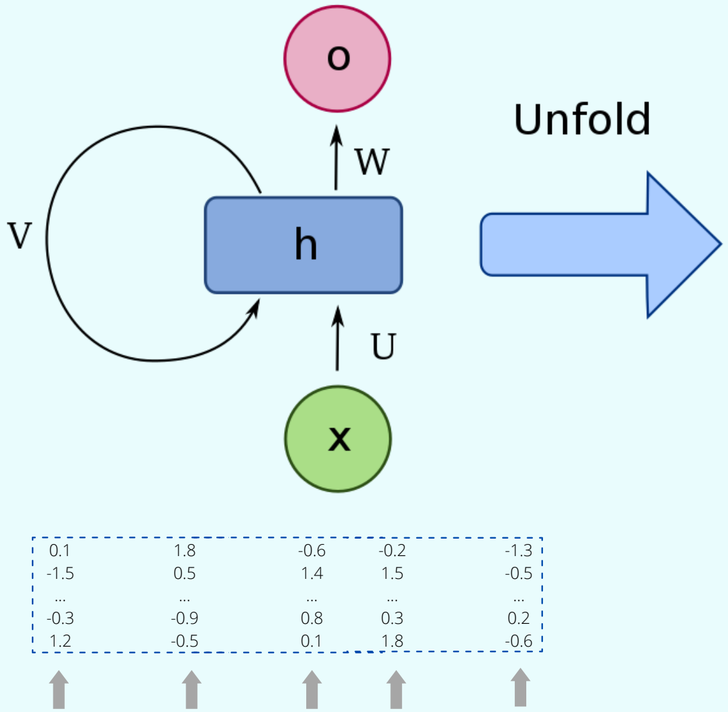
/3 MODELAMIENTO

MODELO RNN

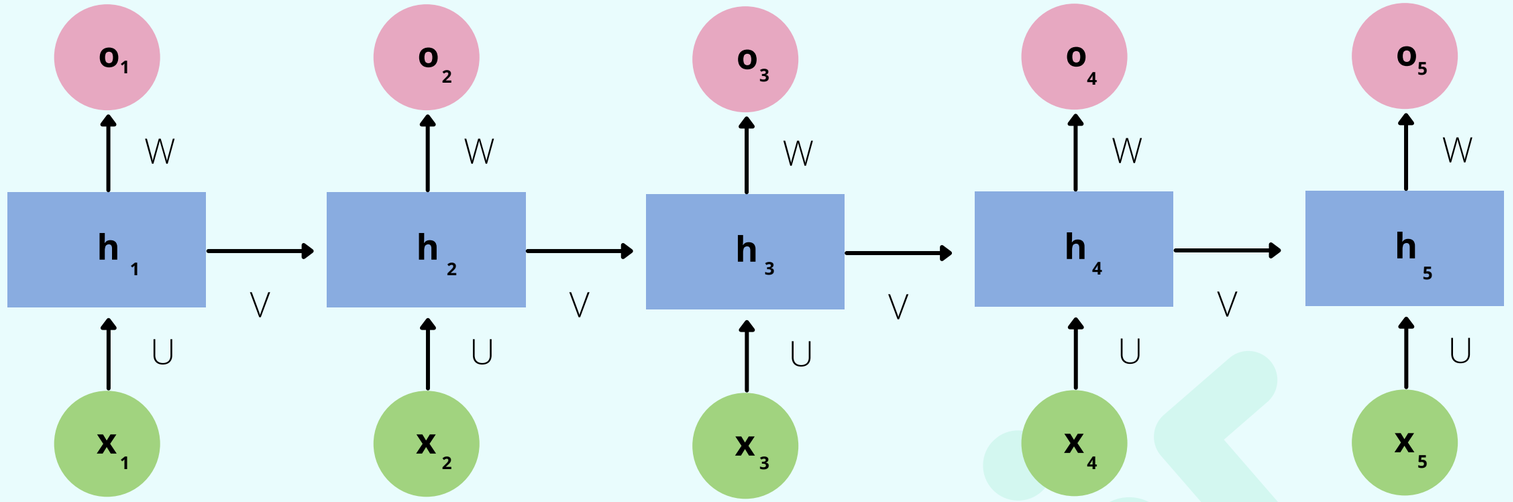
- Recurrent Neural Network (RNN)
- Una capa de embedding, 3 capas de LSTM y una capa lineal.
- A todas se les aplica dropout de 0.5.
- Entrada: vectores one-hot.
- Métricas a utilizar: recall, precision y puntuación F1.

MODELO RNN

PRINCIPIO_ACTIVO | FORMA-FARMA | PERIODICITY | DURATION | ADMIN



RANITIDINA 50 MG SOLUCIÓN INYECTABLE



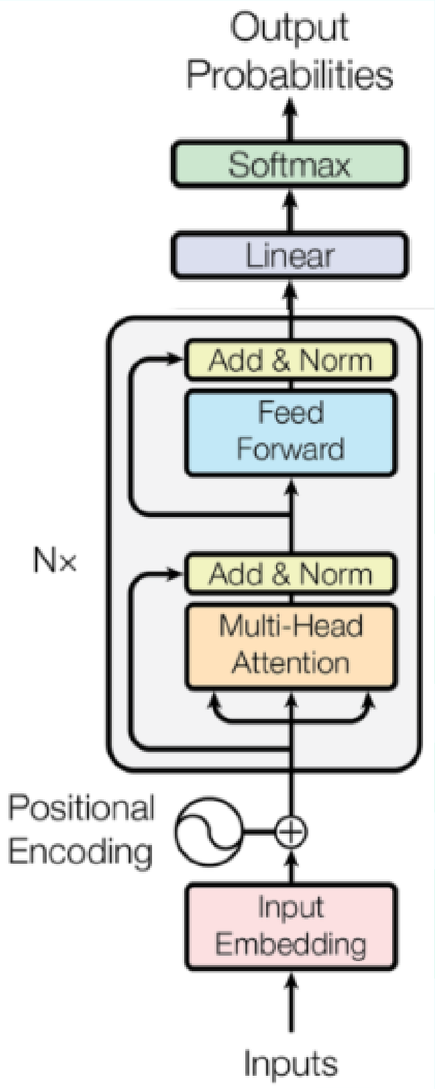
RANITIDINA 50 MG SOLUCIÓN INYECTABLE

MODELO BETO

plncmm / **bert-clinical-scratch-wl-es** like 0

Fill-Mask | PyTorch | Transformers | bert | generated_from_trainer | AutoTrain Compatible

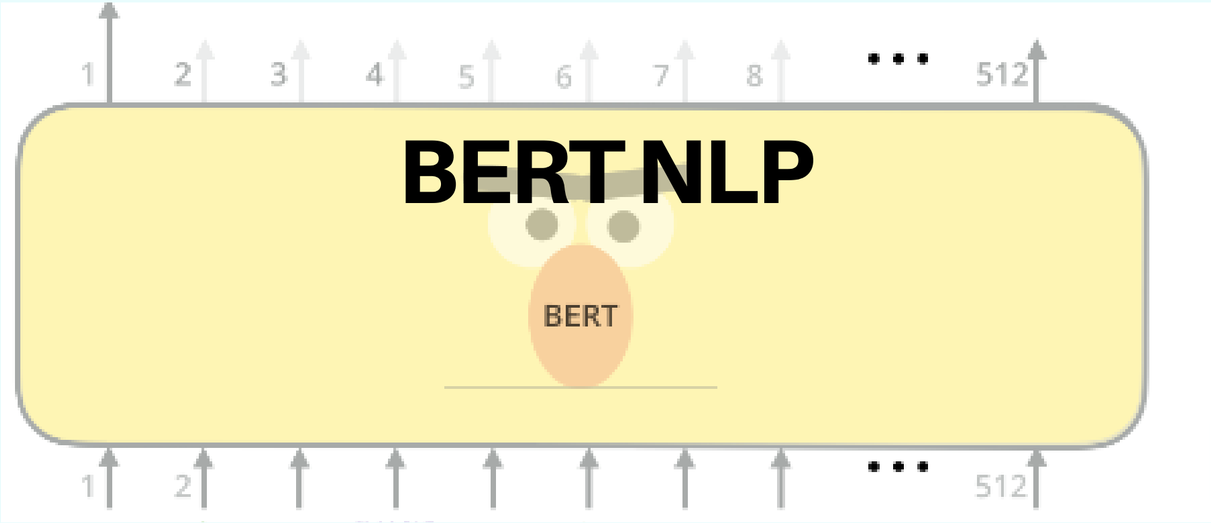
- Modelo basado en Transformers
- Un modelo de 12 capas pre-entrenado
- Fine-tuning con datos clínicos
- Entrada: tokenizador, embeddings iniciales y codificación posicional
- Fine-tuning para entidades
- Métricas a utilizar: recall, precision y puntuación F1.



MODELO BETO

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

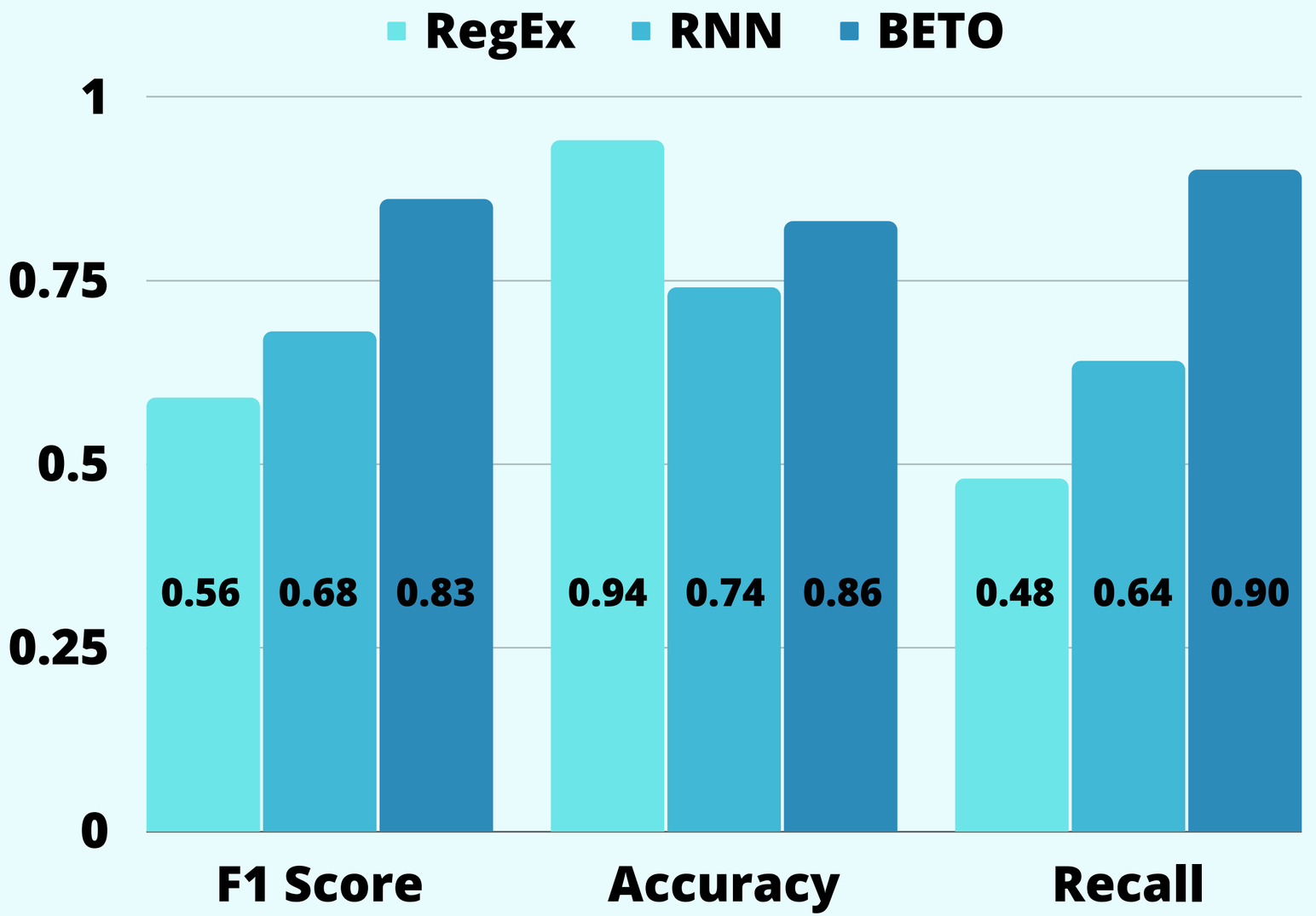
PERIODICITY | DURATION



HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

/4 RESULTADOS

Modelos para 5 Entidades



/5 NUEVAS ITERACIONES.

PRIMEROS MODELOS CON 5 ENTIDADES

Trabajo Paralelo

FINE TUNING

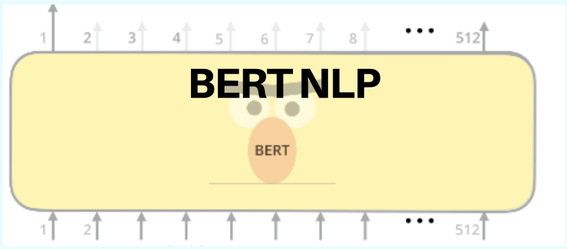
NUEVAS ETIQUETAS

NUEVOS RESULTADOS
Y CONCLUSIONES

MODELO BETO

plncmm / **bert-clinical-scratch-wl-es** like 0

Fill-Mask | PyTorch | Transformers | bert | generated_from_trainer | AutoTrain Compatible



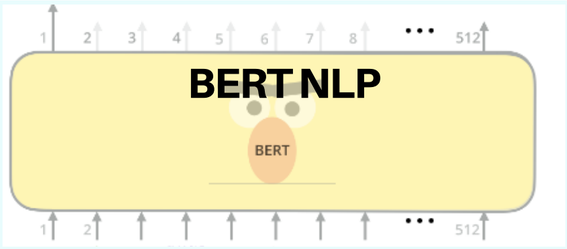
Pre-entrenado con big spanish corpus

Fine-tuning con dataset clínico

MODELO BETO

plncmm / **bert-clinical-scratch-wl-es** like 0

Fill-Mask | PyTorch | Transformers | bert | generated_from_trainer | AutoTrain Compatible



Pre-entrenado con big spanish corpus

Fine-tuning con dataset clínico

MODELO REGEX

Conocimiento experto y lógica

Dataset 100k

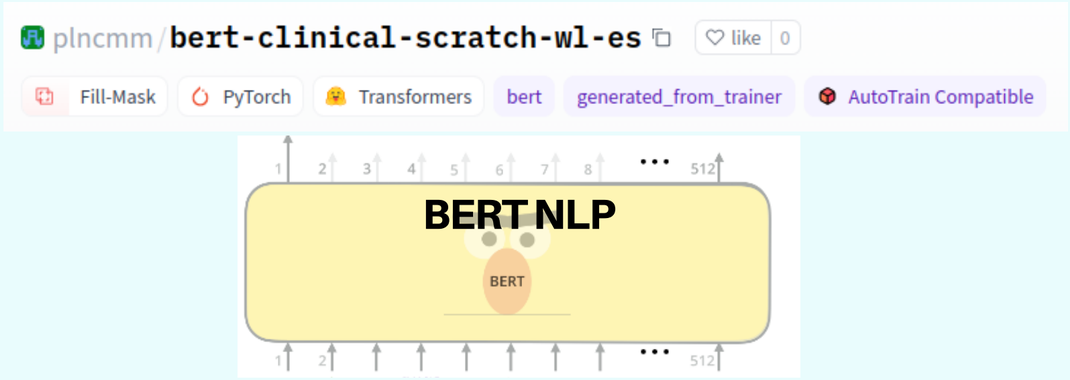
Datos para entrenar (RegEx)

Fine-tuning para reconocimiento de entidades

PRINCIPIO_ACTIVO **FORMA-FARMA**

PERIODICITY **DURATION** **ADMIN**

MODELO BETO



Pre-entrenado con big spanish corpus

Fine-tuning con dataset clínico

MODELO REGEX

Conocimiento experto y lógica

Dataset 100k

Datos para entrenar (RegEx)

Etiquetado manual de datos

Dataset 1k

Ground truth

Fine-tuning para reconocimiento de entidades

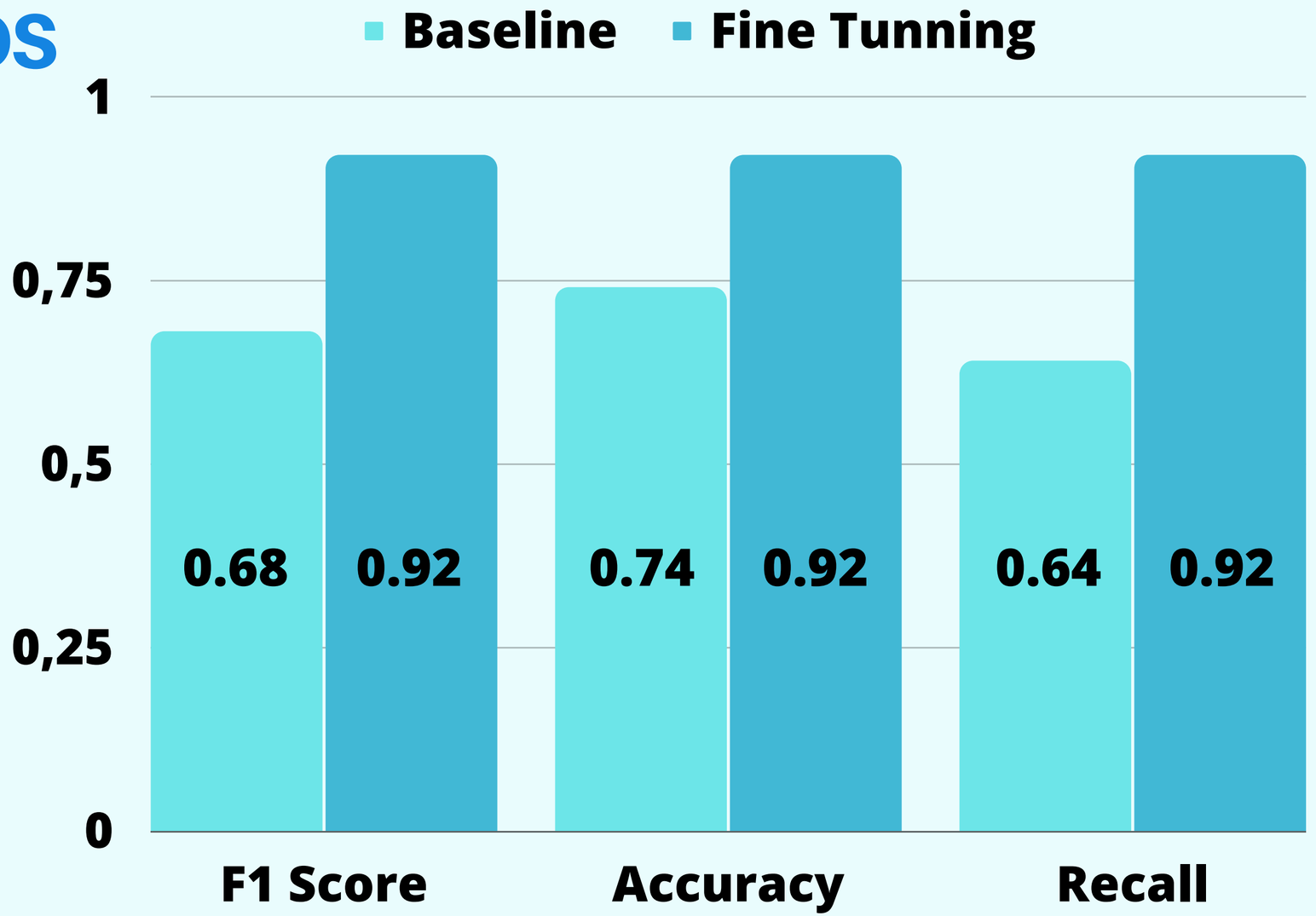
PRINCIPIO_ACTIVO | FORMA-FARMA | PERIODICITY | DURATION | ADMIN

Modelo final

/5 NUEVAS ITERACIONES.

RESULTADOS

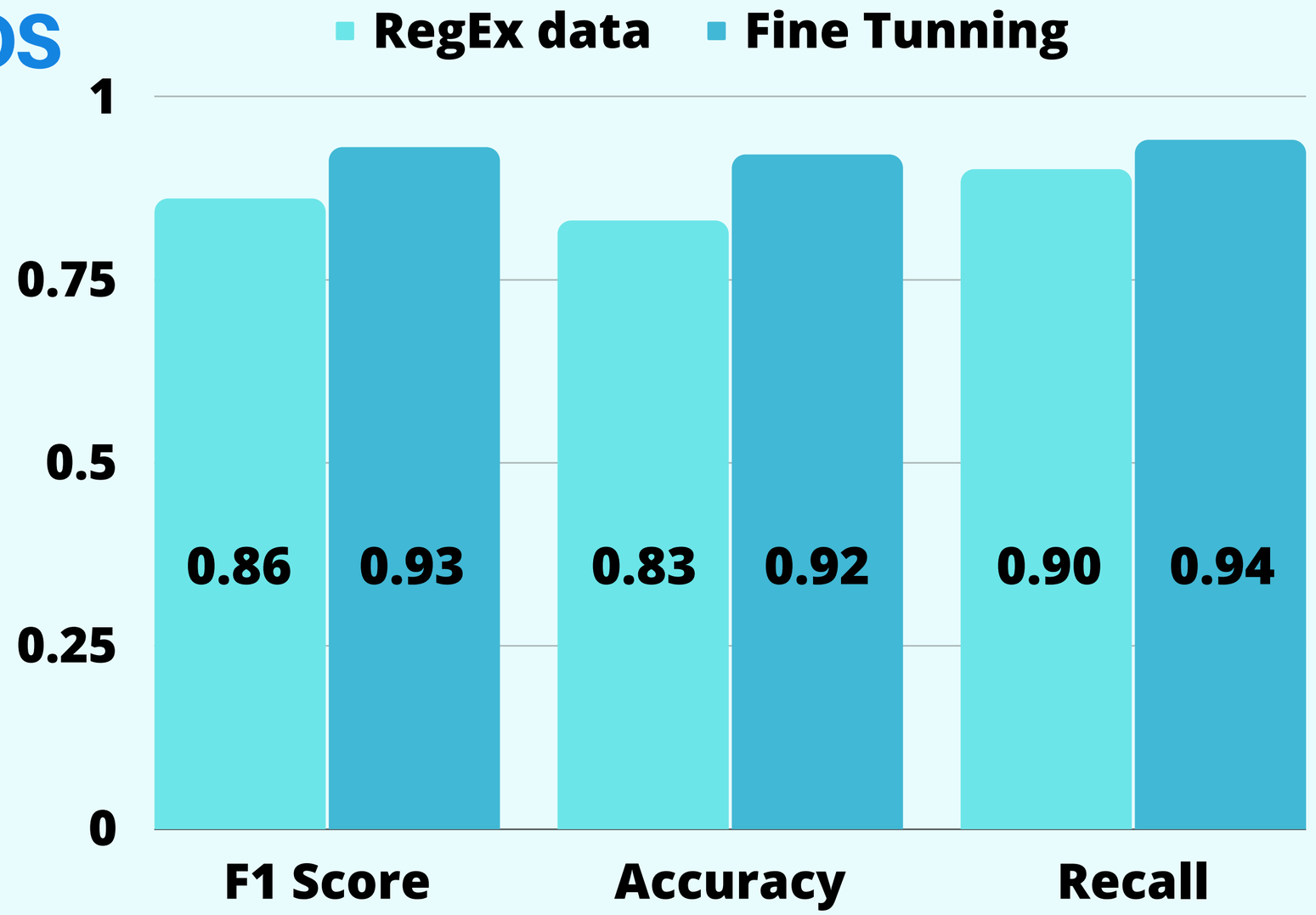
RNN con fine tuning



/5 NUEVAS ITERACIONES.

RESULTADOS

Beto con fine tuning



/5 NUEVAS ITERACIONES: NUEVAS ETIQUETAS

Agregar nuevas etiquetas para obtener más datos relevantes.

ACTIVE_PRINCIPLE | FORMA_FARMA

HIDRALAZINA 50 MG COMPRIMIDO

CANT | UND | VIA_ADMIN

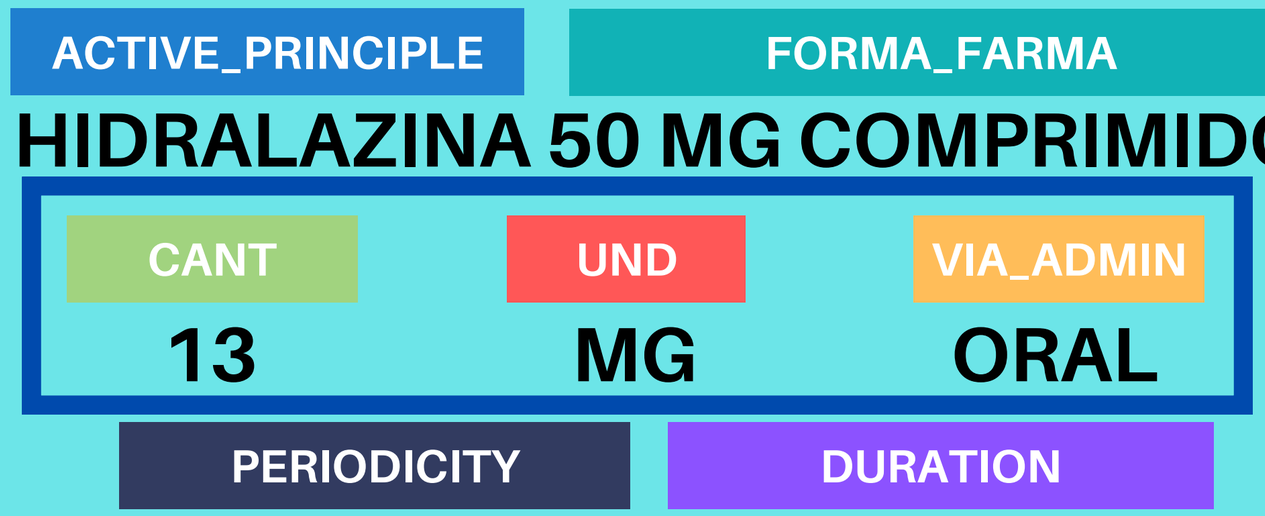
13 **MG** **ORAL**

PERIODICITY | DURATION

cada 12 horas durante 15 dias

/5 NUEVAS ITERACIONES: NUEVAS ETIQUETAS

Agregar nuevas etiquetas para obtener mas datos relevantes.



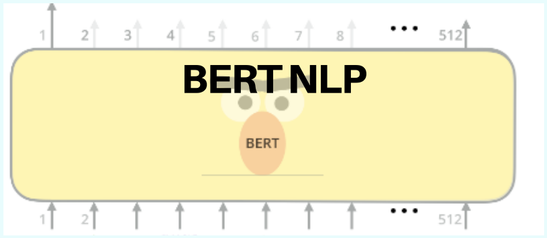
1000 nuevas etiquetas de ADMIN

MODELO BETO ADMIN

copiar modelo

Pre-entrenado con big spanish corpus
Fine-tuning con dataset clínico

Fine-tuning para reconocimiento de entidades
(RegEx + ground truth)



PRINCIPIO_ACTIVO | FORMA-FARMA | PERIODICITY | DURATION | ADMIN

Fine-tuning para reconocimiento de entidades

CANT | UND | VIA_ADMIN

Modelo final

Conocimiento experto y lógica

Etiquetado manual de datos

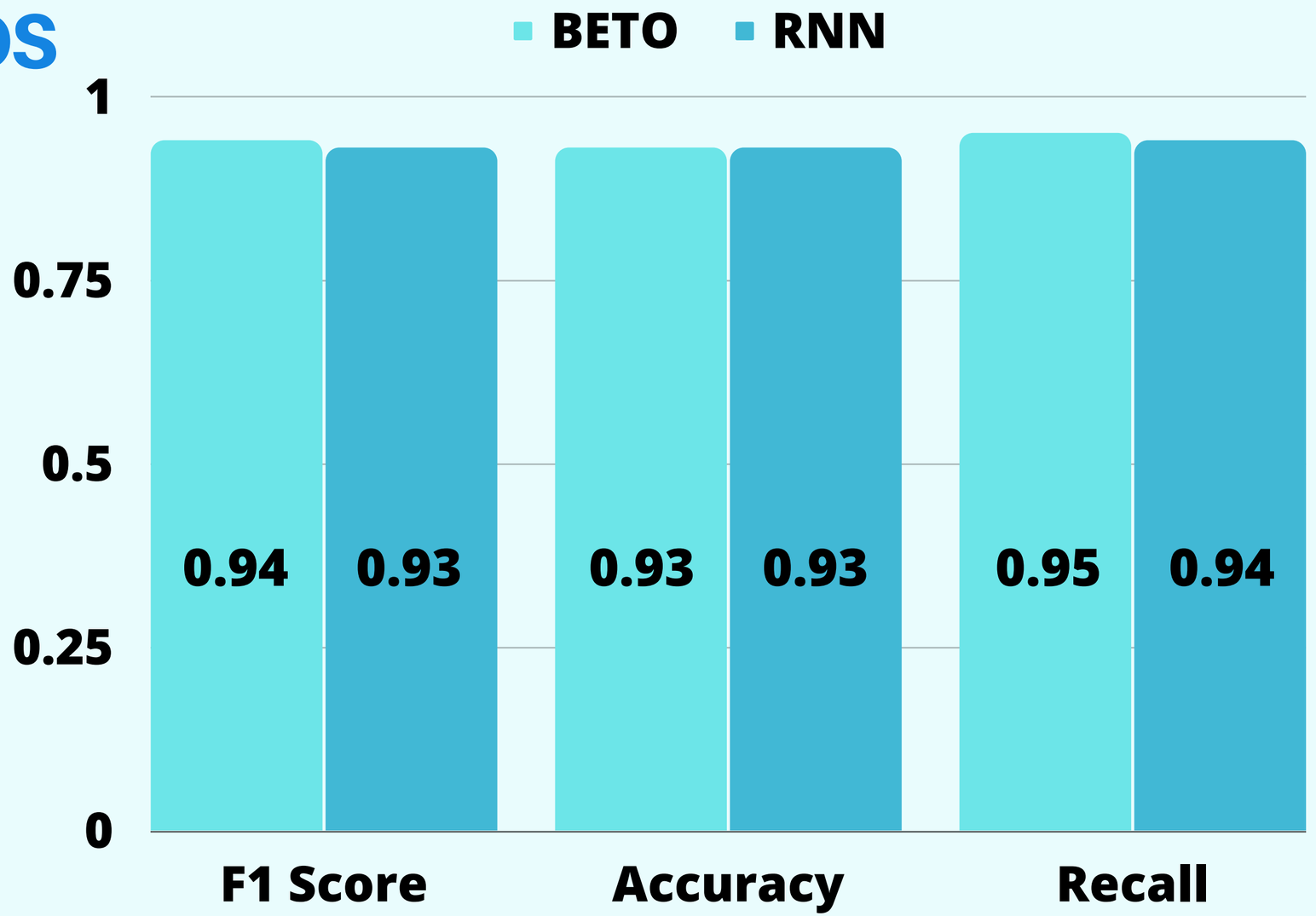
Dataset 1k

Ground truth

/5 NUEVAS ITERACIONES.

RESULTADOS

Modelos para 3 Entidades en ADMIN



/6

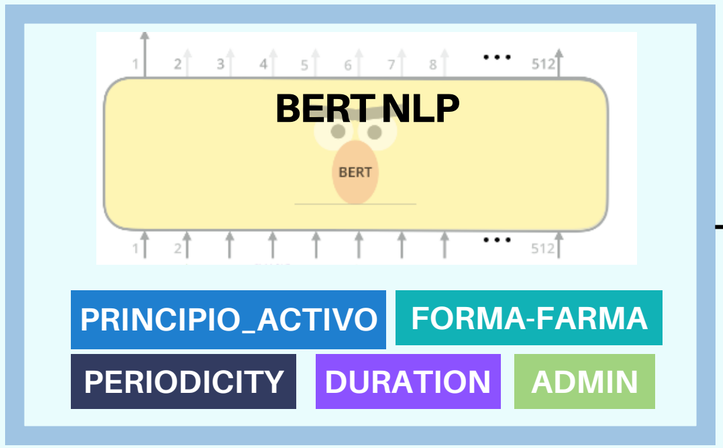
RESUMEN Y RECOMENDACIONES

¿En qué estado están los modelos? ¿Qué se puede lograr con estos? ¿De qué manera pueden aportar?

MODELO BETO ADMIN

Uso de ambos modelos

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS



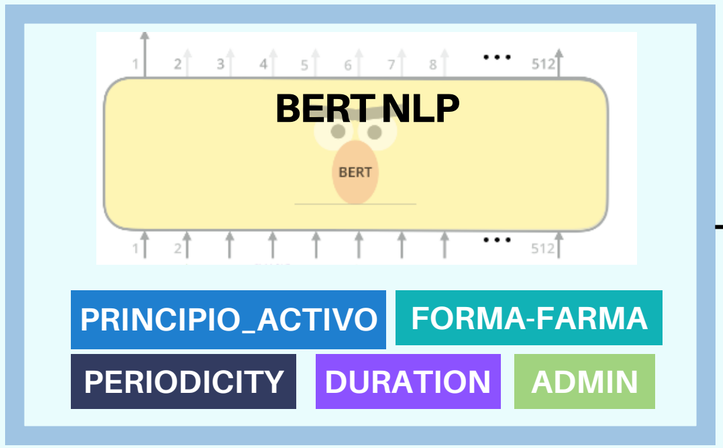
PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN | PERIODICITY | DURATION

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

MODELO BETO ADMIN

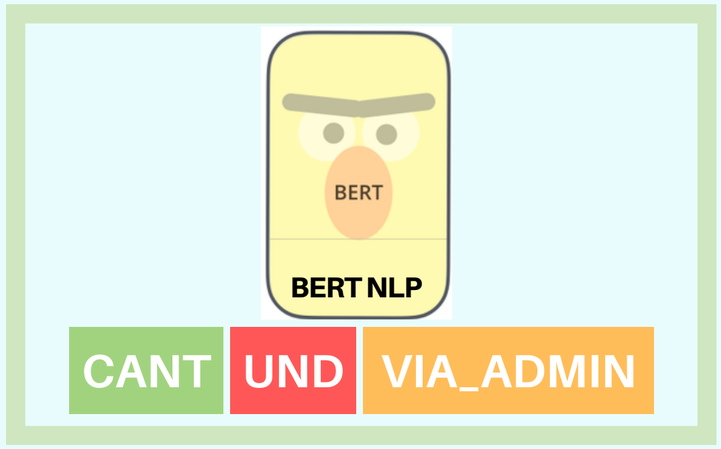
Uso de ambos modelos

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS



PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN | PERIODICITY | DURATION

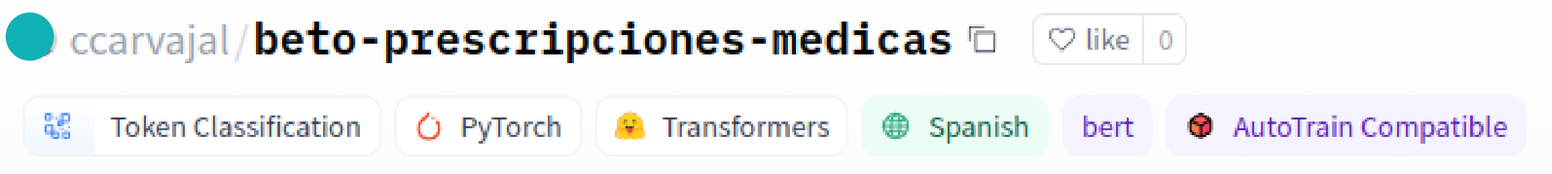
HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

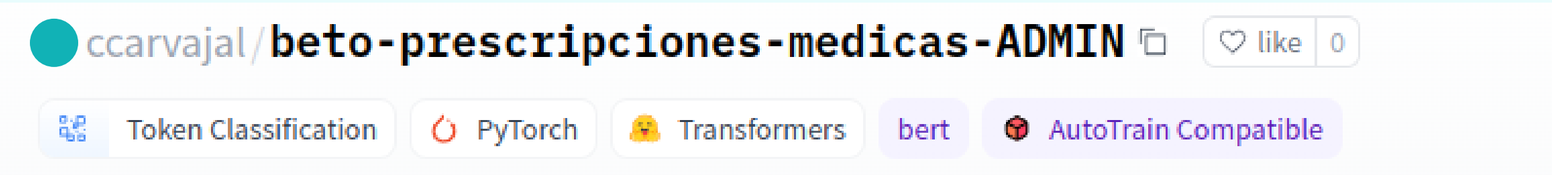


CANT | UND | VIA_ADMIN

DEMO

Modelos disponibles en repositorio HuggingFace





Demo disponible en repositorio Github

```
text = "PARACETAMOL 500 MG COMPRIMIDO 1 COMPRIMIDO ORAL cada 6 horas durante 3 dias"

etiquetar(text)
```

✓ 0.7s Python

ACTIVE_PRINCIPLE	FORMA_FARMA	CANT-ADMIN	UND-ADMIN	VIA-ADMIN	PERIODICITY	DURATION
PARACETAMOL	500 MG COMPRIMIDO	1	COMPRIMIDO	ORAL	cada 6 horas	durante 3 dias

DETECCIÓN DE ERRORES

Input

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

DETECCIÓN DE ERRORES

Input

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

Output modelo

PRINCIPIO_ACTIVO	FORMA-FARMA	ADMIN	PERIODICITY	DURATION
HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS				

CANT	UND	VIA_ADMIN

DETECCIÓN DE ERRORES

Input

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

Output modelo

PRINCIPIO_ACTIVO	FORMA-FARMA	ADMIN	PERIODICITY	DURATION

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

CANT | UND | VIA_ADMIN

Conocimiento experto y lógica

DETECCIÓN DE ERRORES

Input

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

Output modelo

PRINCIPIO_ACTIVO	FORMA-FARMA	ADMIN	PERIODICITY	DURATION

HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL CADA 12 HORAS DURANTE 15 DIAS

CANT | UND | VIA_ADMIN

Conocimiento experto y lógica

Alertas en caso de errores de prescripción

/6 RECOMENDACIONES

Agregar más etiquetas para obtener mas datos relevantes.

ACTIVE_PRINCIPLE FORMA_FARMA

HIDRALAZINA 50 MG COMPRIMIDO

CANT UND VIA_ADMIN

13 MG ORAL

PERIODICITY DURATION

cada 12 horas durante 15 dias

/6 RECOMENDACIONES

Agregar más etiquetas para obtener mas datos relevantes.

ACTIVE_PRINCIPLE | FORMA_FARMA

HIDRALAZINA 50 MG COMPRIMIDO

CANT | UND | VIA_ADMIN

13 MG ORAL

ART	CANT	TIEMPO	ART	CANT	TIEMPO
cada	12	horas	durante	15	dias

/6 CONCLUSIONES

- El lenguaje semi-estructurado influyó en la obtención de buenas métricas.
- Los modelos se diferencian en gasto computacional y reproducibilidad.
- Beto posee mas reproducibilidad y robustez, se debe comprobar con datos de otros hospitales.
- El estado actual de los modelos puede detectar la gran mayoría de errores de incompletitud.
- Da pie a la posibilidad de agrupar y generar mayor cantidad de datos para el uso de otros algoritmos de detección de outliers.

MDS Master of Data Science
Universidad de Chile

PRESENTACIÓN FINAL
ENTIDADES MINSAL

DANIEL CARMONA, MARTÍN SEPÚLVEDA,
MONSERRAT PRADO, CAMILO CARVAJAL

REFERENCIAS

- Sang, E. F., De Meulder, F.

Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.
In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (142–147), 2003.

- Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., Ghosh, P.

A survey on recent named entity recognition and relationship extraction techniques on clinical texts.
In Applied Sciences (11(18), 8319.), 2021.

- Báez, P., Villena, F., Rojas, M., Durán, M., Dunstan, J. (2020, November).

The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish.
InProceedings of the 3rd clinical natural language processing workshop (pp. 291-300)., 2020.

- Báez, P., Bravo-Marquez, F., Dunstan, J., Rojas, M., Villena, F.

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish.
In ACM Transactions on Computing for Healthcare, (3(3), 1-22.) - 2022.

- Rojas, M., Dunstan, J., Villena, F.

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing.
In Proceedings of the 4th Clinical Natural Language Processing Workshop, (pp. 87-92)., 2022.

- Jiang, M., Sanger, T., Liu, X.

Combining contextualized embeddings and prior knowledge for clinical named entity recognition: evaluation study.
In JMIR medical informatics, (7(4), e14850.) - 2019