MDS Master of Data Science Universidad de Chile

PRESENTACIÓN 4 MDS7201

SOLUCIÓN AL PROBLEMA Y RESULTADOS

ENTIDADES MINSAL

DANIEL CARMONA, MARTÍN SEPÚLVEDA,
MONSERRAT PRADO, CAMILO CARVAJAL

ENTIDADES MINSAL

TABLA DE CONTENIDO

/1 RESUMEN

/2 MODELAMIENTO

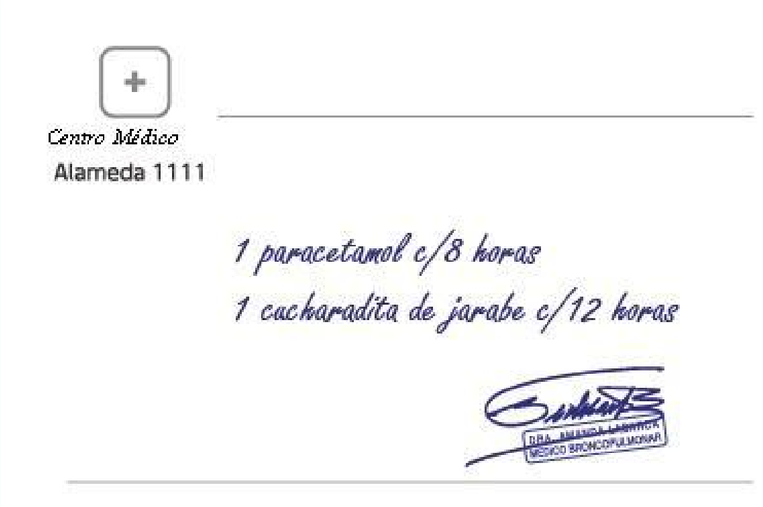
/3 RESULTADOS

/4 EVALUACIÓN, ANÁLISIS Y CONCLUSIONES

/5 TRABAJO FUTURO

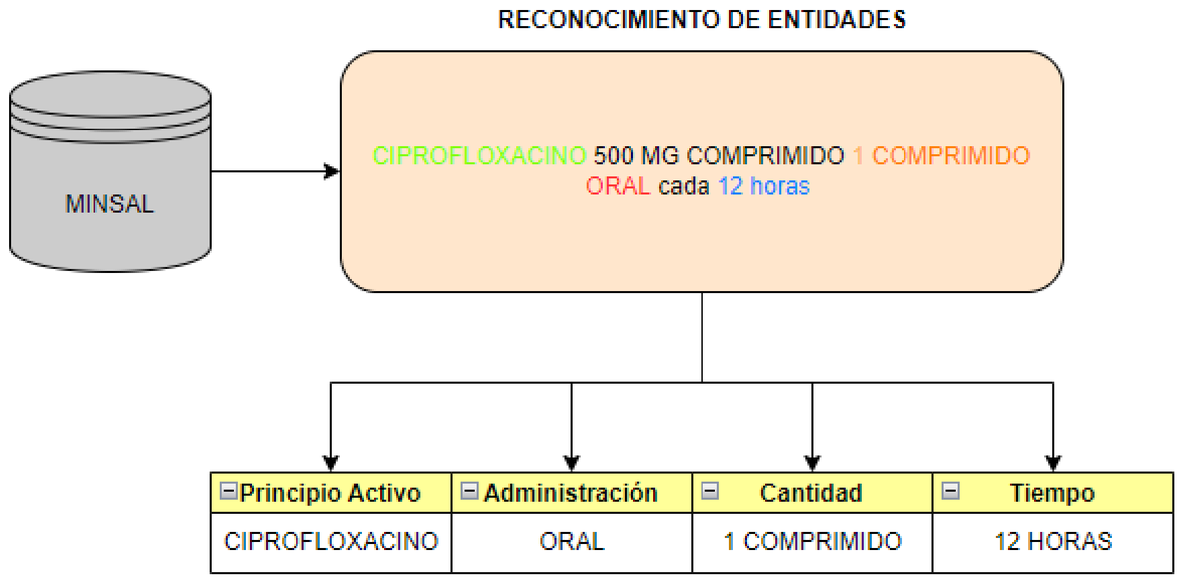
/1 RESUMEN

- 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.



RESUMEN
DESCRIPCIÓN DEL PROYECTO

- Dado un campo de texto libre, utilizar algoritmos de NLP para reconocer entidades y completar columnas de manera automática en los datos de un paciente.
- Detectar errores de completitud o gramática en las indicaciones.
- Refraseo de la información para evitar errores de administración de medicamentos.



RESUMEN DATOS

- 1.5 [M] de prescripciones, con un total de 20 atributos por cada una

	CODIGO_MEDICAMENTO	PRES_DENOMINACION	RESUMEN	IND_ADMINISTRACION_1	IND_ADMINISTRACION_2
1526553	FACC09001	CAPTOPRIL 25 MG COMPRIMIDO	1 COMPRIMIDO ORAL cada 8 horas	NaN	NaN
1526554	FANN02016	PARACETAMOL 500 MG COMPRIMIDO	2 COMPRIMIDO ORAL cada 8 horas	NaN	NaN
1526555	FAAA10002	INSULINA CRISTALINA HUMANA 100 U.I./ML SOLUCIO...	2 UNIDAD INTRAVENOSA cada 6 horas	NaN	NaN

- En ciertos atributos se cuenta con un gran porcentaje de valores vacíos o NaN.
- Estos no se consideran relevantes para el entrenamiento del modelo.

RESUMEN LITERATURA

NER: Named Entity Recognition

Sang | Meulder — 2003 — ↳ 2419

Introduction to the CoNLL-2003 shared task: language-independent named entity recognition

CoNLL

Texto Clínico

Bose | ... | Ghosh — 2021 — ↳ 4

A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts

Applied Sciences

Báez | ... | Villena — 2022 — ↳ 0

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish

ACM transactions on computing for healthcare

Báez | ... | Dunstan — 2020 — ↳

The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish

Association for Computational Linguistics

Dunstan | Villena — 2022 — ↳ 0

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing

Association for Computational Linguistics

Contexto Chileno

RESUMEN LITERATURA

Texto Clínico

Báez ... Dunstan — 2020

The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish

Association for Computational Linguistics

Báez ... Villena — 2022 — ↳ 0

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish

ACM transactions on computing for healthcare

Dunstan Villena — 2022 — ↳ 0

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing

Association for Computational Linguistics

Conocimiento previo

Jiang ... Liu — 2019

Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study.

JMIR medical informatics

Akbik ... Vollgraf — 2019

FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP

Association for Computational Linguistics

Kazama Torisawa — 2007 — ↳ 266

Exploiting Wikipedia as External Knowledge for Named Entity Recognition

EMNLP

Modelos de lenguaje

Devlin ... Toutanova — 2019

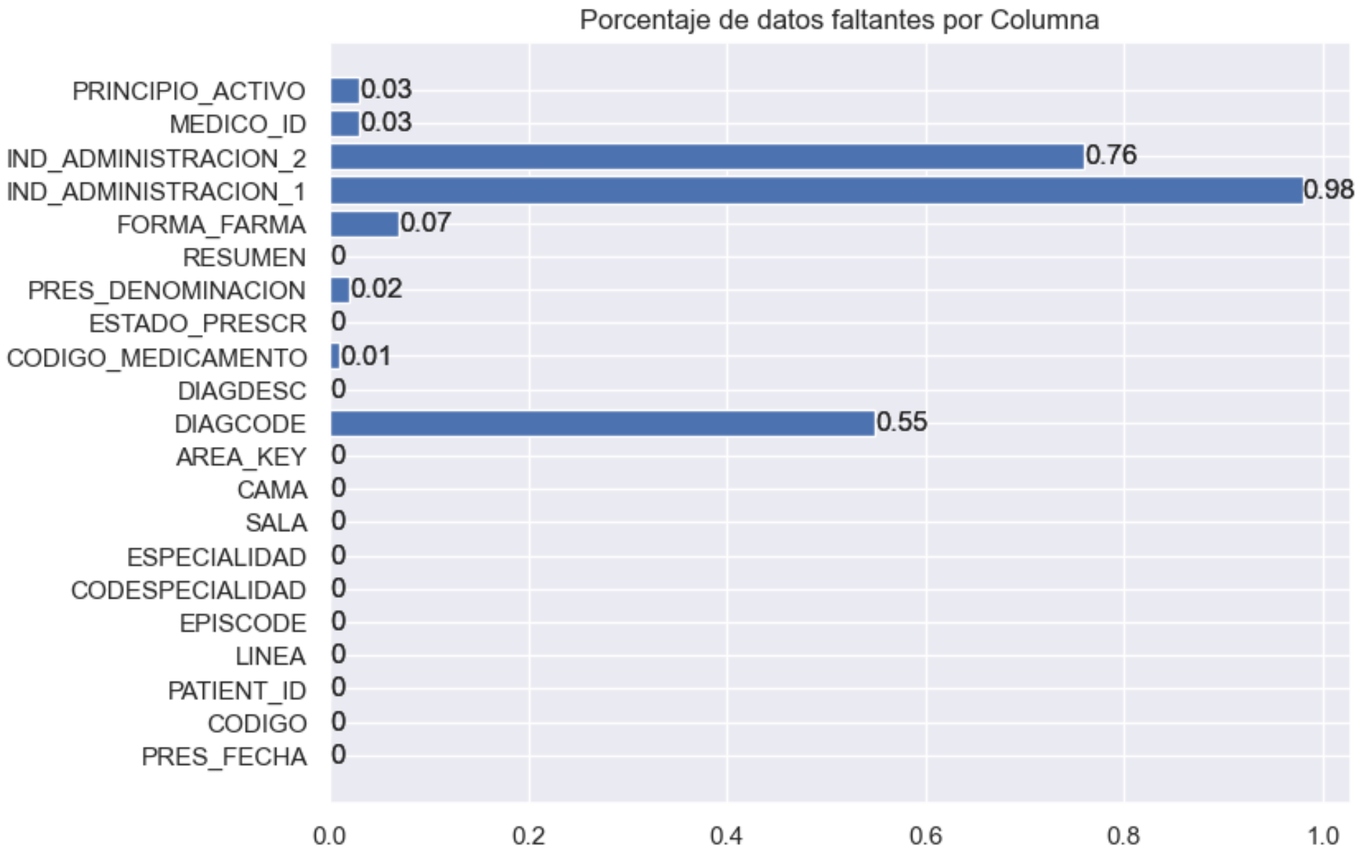
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Association for Computational Linguistics

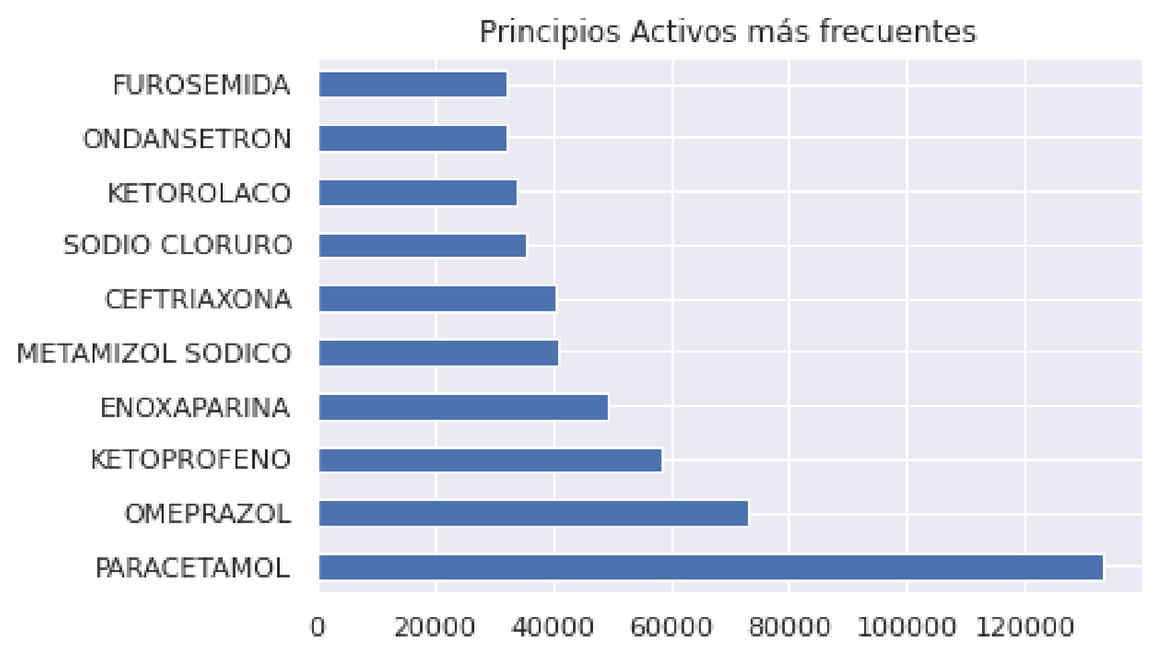
DATA FALTANTE Y ADICIONAL

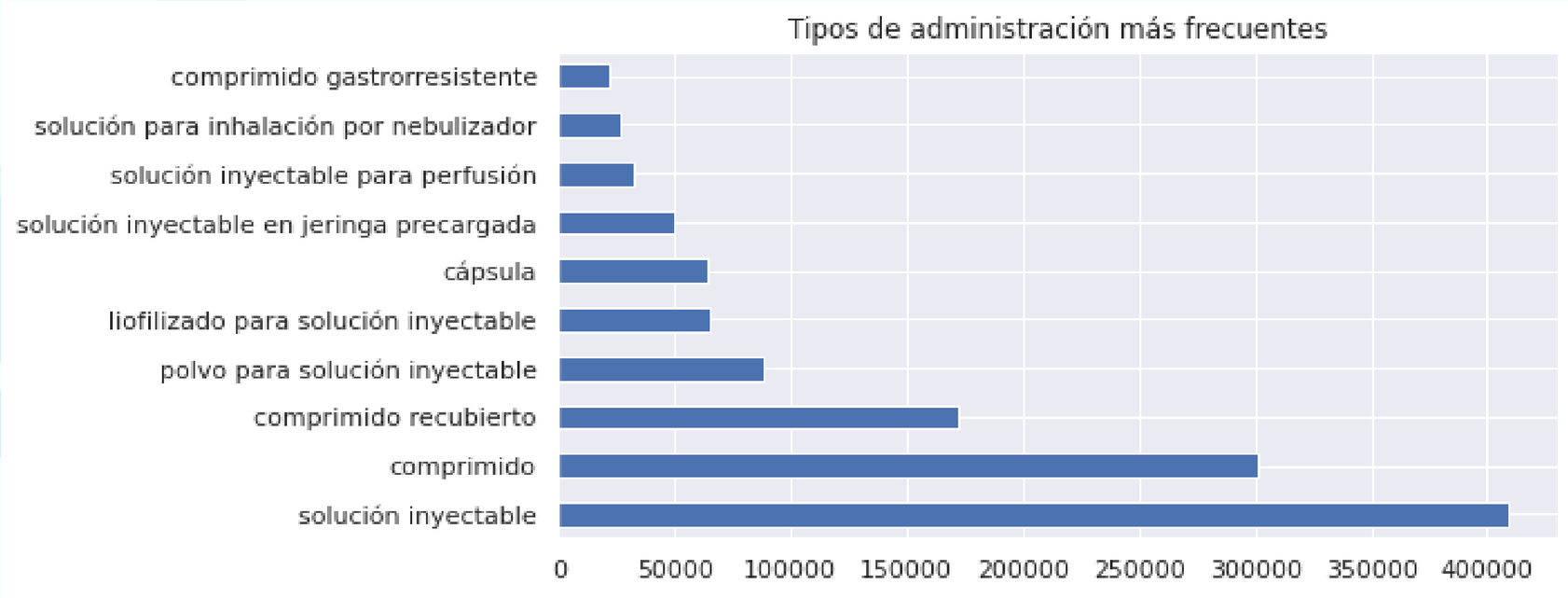
- Se agregó una nueva columna de Principios activos siguiendo el código HLF.
- Gran parte de los datos faltantes corresponden a los atributos Indicación de Administración 1 y 2, los cuales son utilizados para casos especiales de administración.
- Datos faltantes en códigos de medicamentos: **18379**
- Datos faltantes en Principio activo: **50437**

No todos los códigos de medicamento en las prescripciones tienen código HLF asociado.



VISUALIZACIÓN DE DATOS



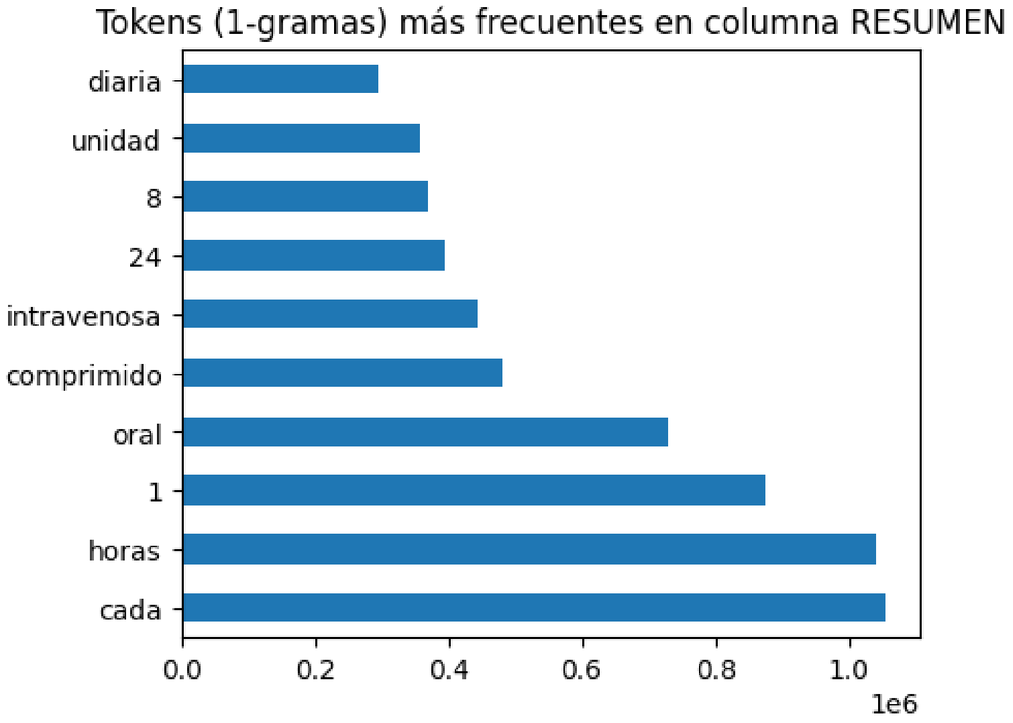


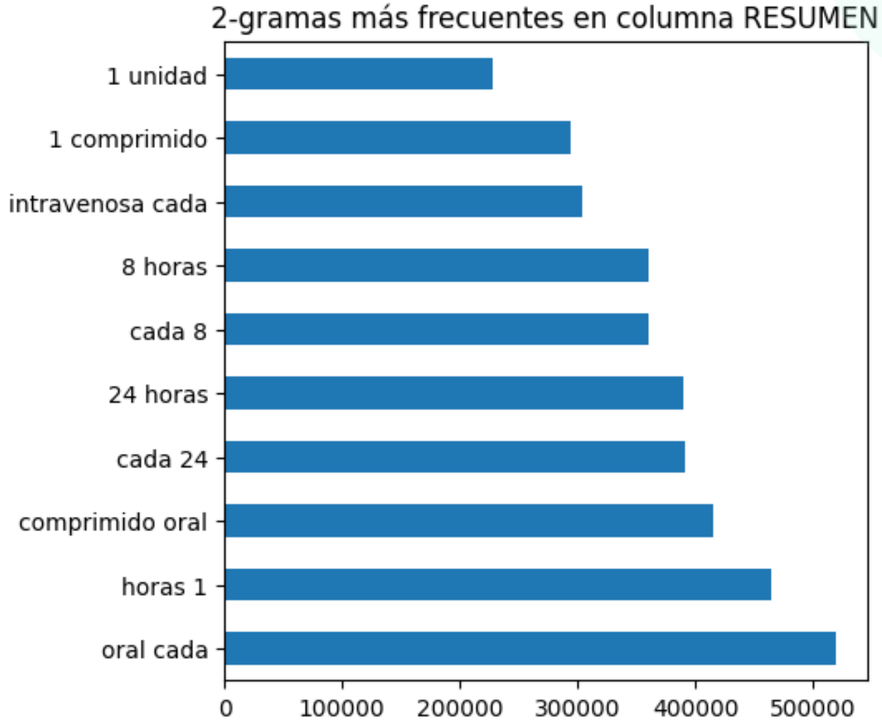
PALABRAS COMUNES EN TEXTO LIBRE: RESUMEN DE LA PRESCRIPCIÓN

Cantidad de valores nulos en columna: 1

Total de 1473779 filas duplicadas (96.5%)

Cantidad de valores únicos : 52777 (3.5%)





PLANTEAMIENTO DE SOLUCIÓN

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

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

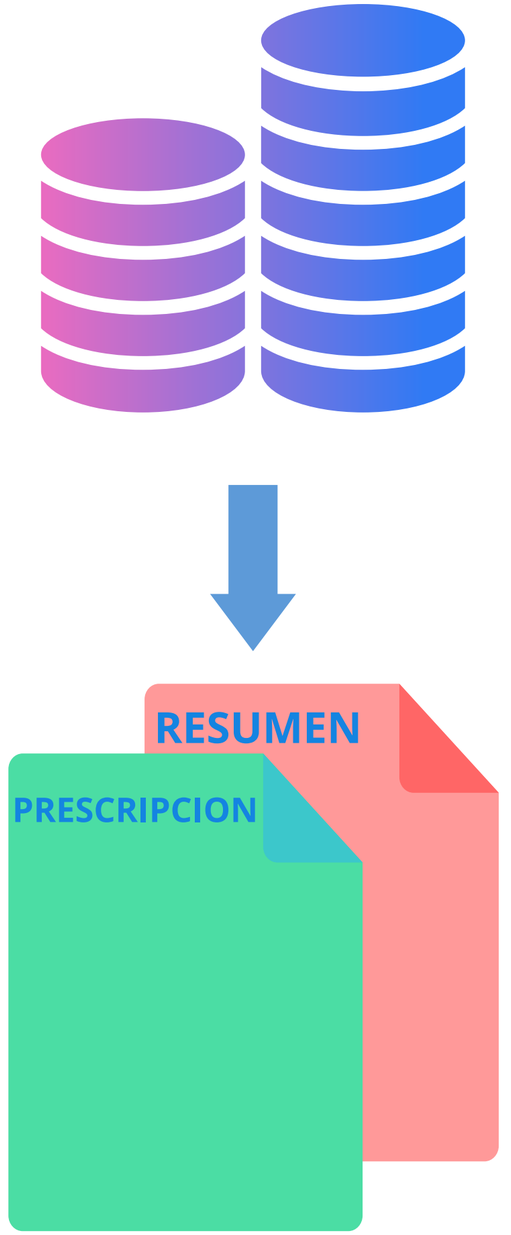
HIDRALAZINA 50 MG COMPRIMIDO 13 MG ORAL

PERIODICITY | DURATION

cada 12 horas durante 15 dias

PREPARACIÓN DE LOS DATOS

- Se define el texto libre como las columnas: PRES_DENOMINACION y RESUMEN.
- Otras columnas útiles PRINCIPIO_ACTIVO y FORMA_FARMA
- Número de Ejemplos únicos etiquetados se reduce a 108.049



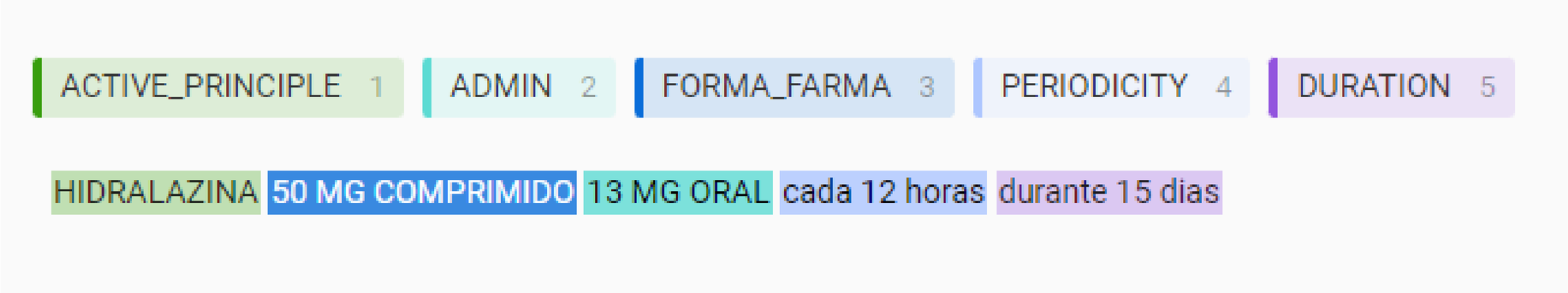
PREPARACIÓN DE LOS DATOS

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

PARACETAMOL	B-ACTVPRNCP
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

ETIQUETADO DE DATOS MANUAL

- Se utilizó la herramienta Label Studio.
- Se etiquetaron 1000 recetas.
- Cada integrante etiquetó 250 datos.



MODELO REGEX

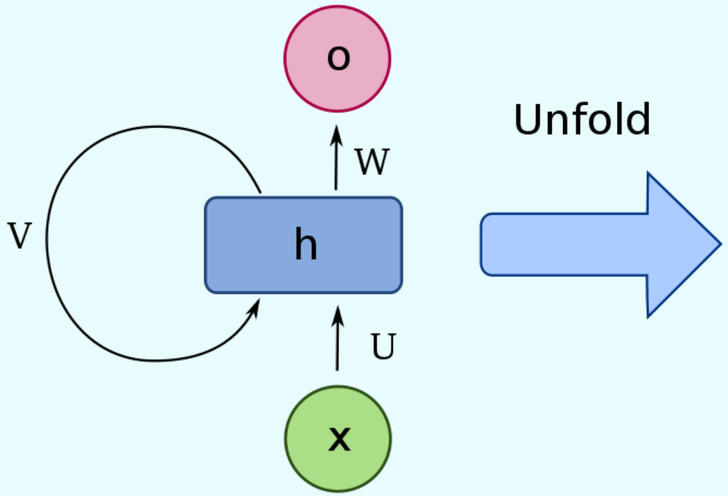
- Se utlizan 2 conjuntos de Principio Activo y Forma Farma
- Se reconocen expresiones regulares de Periodicidad , Duración y Admin.

```
['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-ADMIN']
['CADA'         'B-PERIODICITY']
['8'            'I-PERIODICITY']
['HORAS'        'I-PERIODICITY']
['DURANTE'      'B-DURATION']
['15'           'I-DURATION']
['DIAS'         'I-DURATION']
```

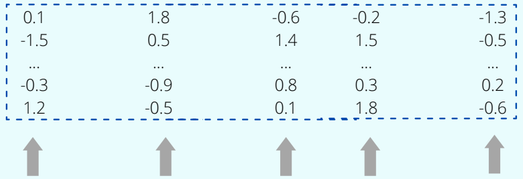
MODELO RNN

- 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



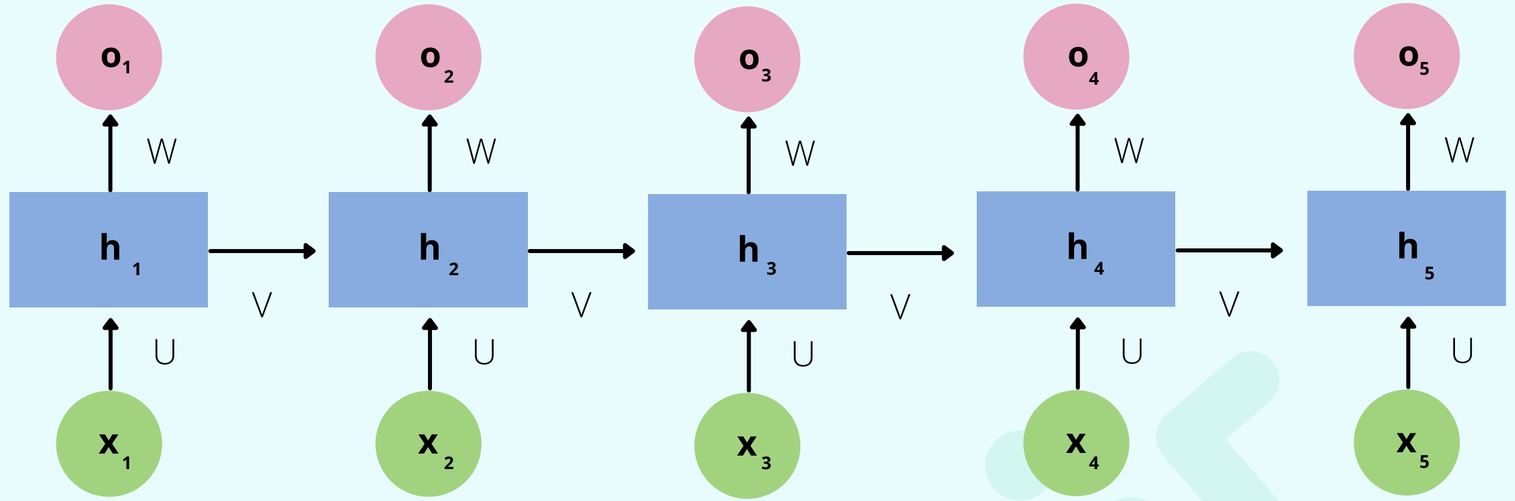
Unfold



RANITIDINA 50 MG SOLUCIÓN INYECTABLE

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

PERIODICITY | DURATION



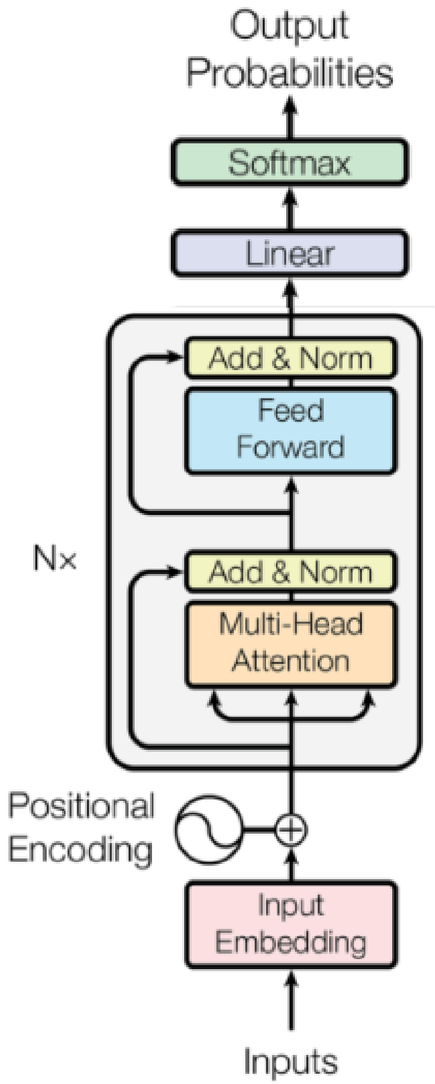
RANITIDINA 50 MG SOLUCIÓN INYECTABLE

MODELO BETO

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

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

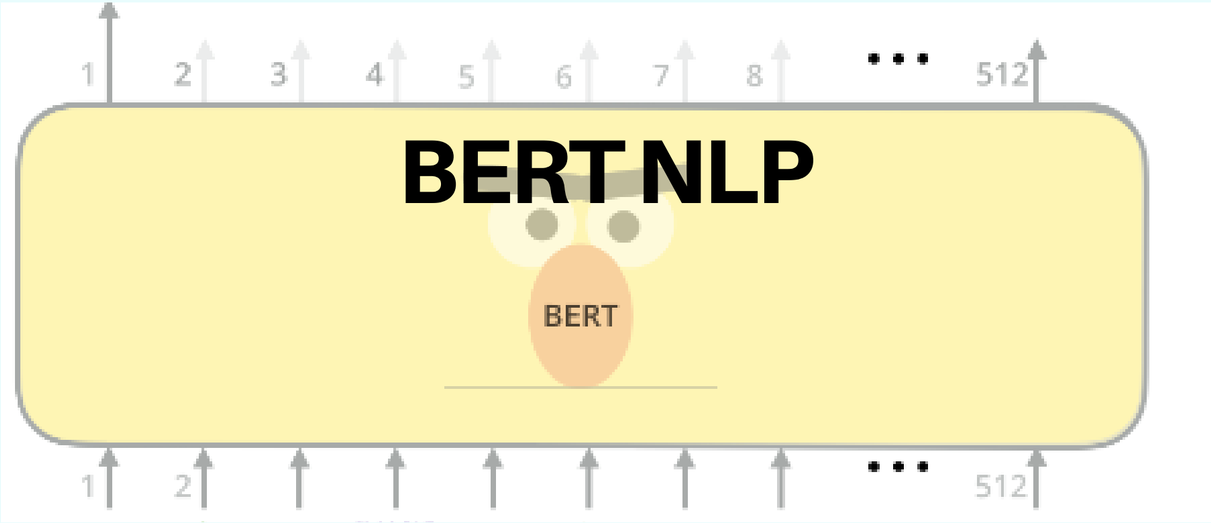
- Modelo basado en Transformers
- Un modelo de 12 capas pre-entrenado
- Fine-tunning con datos clínicos
- Entrada: tokenizador, embeddings iniciales y codificación posicional
- Fine-tunning para entidades
- Métricas a utilizar: recall, precision y puntuación F1.



MODELO BETO

PRINCIPIO_ACTIVO | FORMA-FARMA | ADMIN

PERIODICITY | DURATION



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

/4 RESULTADOS

Modelo	F1 Score	Precision	Recall
RNN - Test (Expresiones regulares)	89%	96%	83%
RegEx - Test (Expresiones regulares)	59%	94%	48%
RNN - Test (Etiquetados Manualmente)	68%	74%	64%
BETO - Test (Etiquetados Manualmente)	75%	68%	82%

Hablando con Expertos

Se requiere de mayor precisión en el campo de Administración, por lo que se separa en Cantidad, Unidad de Medida y Vía de Administración.

PRINCIPIO_ACTIVO | FORMA-FARMA

HIDRALAZINA 50 MG COMPRIMIDO

CANTIDAD | UNIDAD MEDIDA | VIA ADMIN

13 **MG** **ORAL**

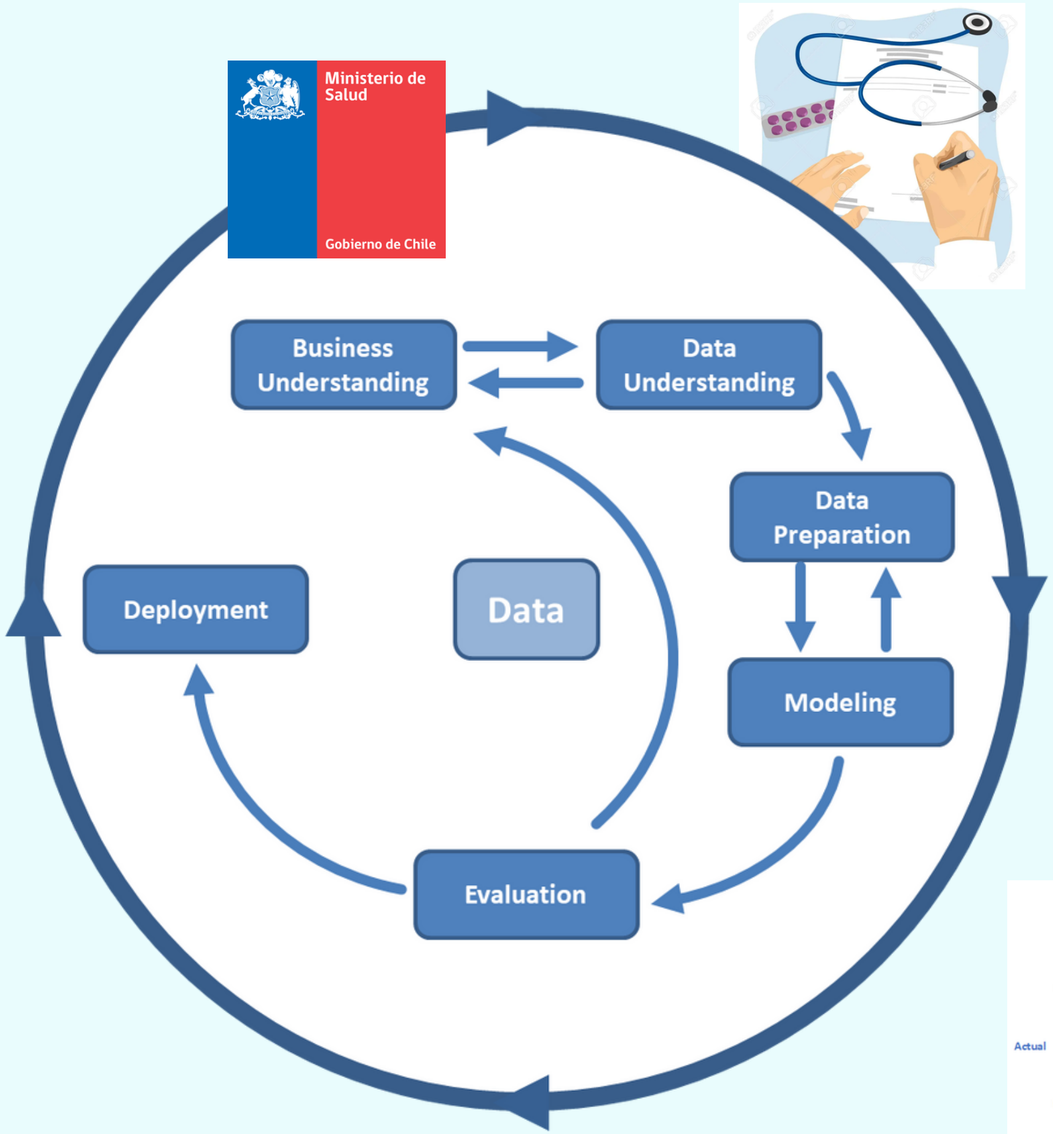
PERIODICITY | DURATION

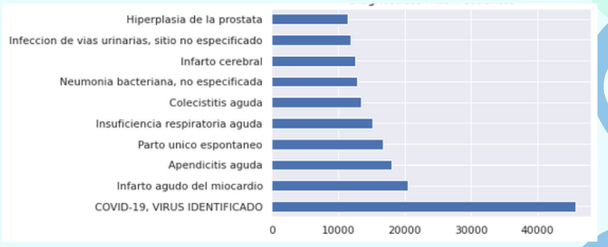
cada 12 horas durante 15 dias

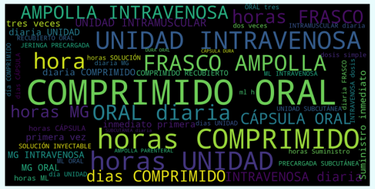
/4 CONCLUSIONES Y TRABAJO FUTURO

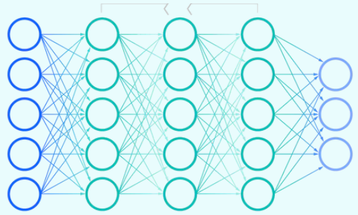
- Fue posible obtener resultados reales desde datos manualmente etiquetados y obtenidos de expresiones regulares.
- En general modelos con redes logran mejores métricas que Expresiones regulares
- Es necesario ejecutar iteraciones mejorando hiperparamentros de las Redes usadas y agregar estrategias como fine tunning y cross validation.
- Mejorar la especificidad de las etiquetas para observar su comportamiento en las Redes

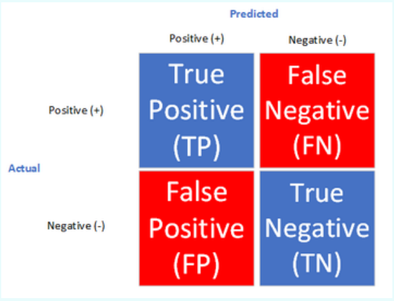
/5 PLAN DE TRABAJO

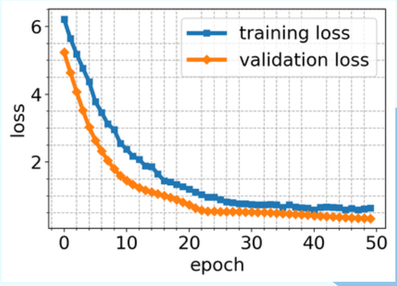












MDS Master of Data Science Universidad de Chile

PRESENTACIÓN 3 MDS7201

PROYECTO DEFINIDO

ENTIDADES MINSAL

DANIEL CARMONA, MARTÍN SEPÚLVEDA,
MONSERRAT PRADO, CAMILO CARVAJAL

/5 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