

# PROCESAMIENTO DE CONTRATOS SOCIETARIOS

TRABAJO FINAL DE LA CARRERA DE  
ESPECIALIZACIÓN EN INTELIGENCIA ARTIFICIAL



**AUTOR:**  
**Mg. Ing. Ezequiel Guinsburg**

**DIRECTOR:**  
**Dr. Luciano Del Corro (Microsoft Research)**

**JURADOS:**  
Dr. Ing. María De Los Milagros Gutiérrez (UTN-FRSF)  
Dr. Ing. Lucila Romero (UNL)  
Ing. Juan Esteban Carrique (UNL)

**CLIENTE:**  
Mg. Lic. Diego González (MERCADOLIBRE)

# AGENDA

01

INTRODUCCIÓN

Contexto y motivación

02

VALIDACIÓN DE USUARIOS

Objetivos, desafíos y planteo general del sistema

03

PROCESAMIENTO DE IMÁGENES

Preprocesamiento, *Bounding Box*, OCR

04

PROC. DE LENGUAJE NATURAL

NER (*Named Entity Recognition*)

05

CLASIFICACIÓN BINARIA

Algoritmo y diagrama de flujo

06

IMPLEMENTACIÓN

Arquitectura, video y resultados

07

EPÍLOGO

Conclusiones y trabajos futuros



01

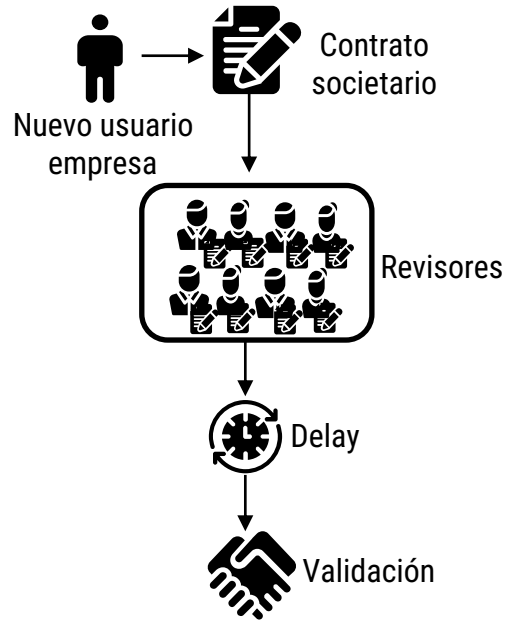
# INTRODUCCIÓN

---

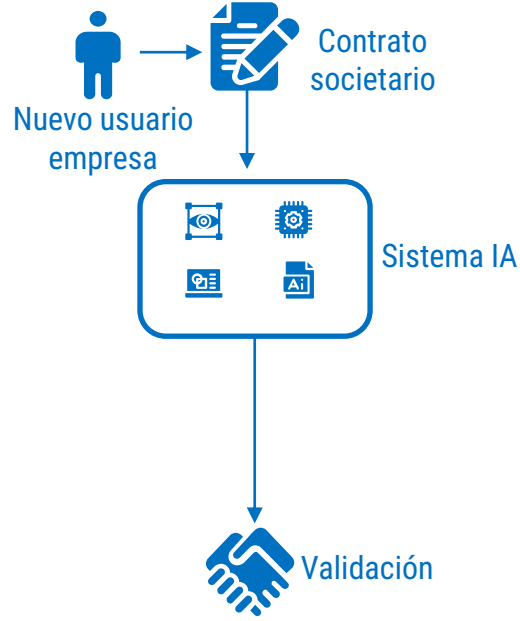
Contexto y motivación



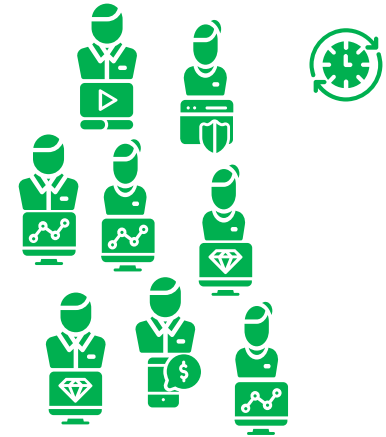
# SITUACIÓN ACTUAL



# SITUACIÓN OBJETIVO



# RESULTADOS

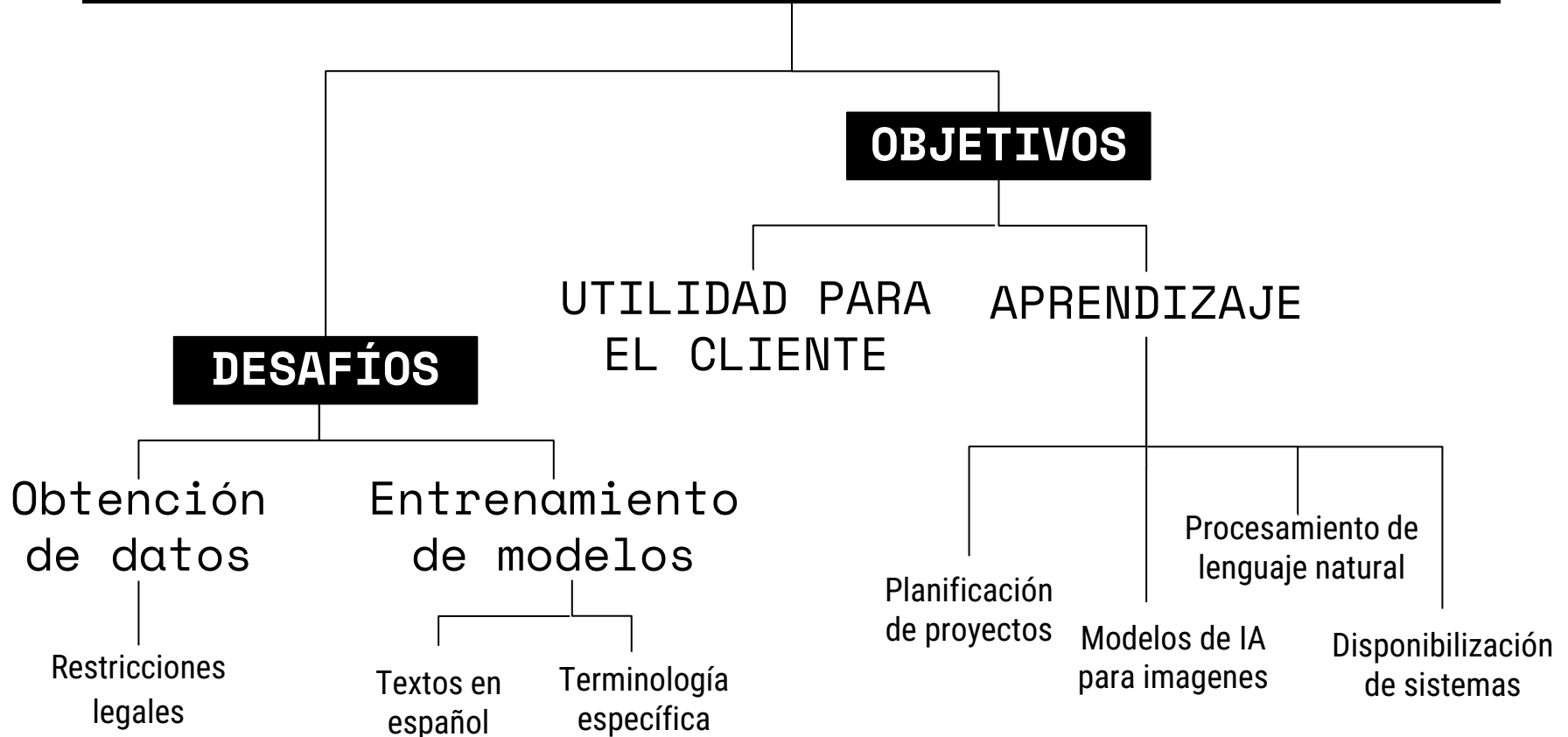


# 02

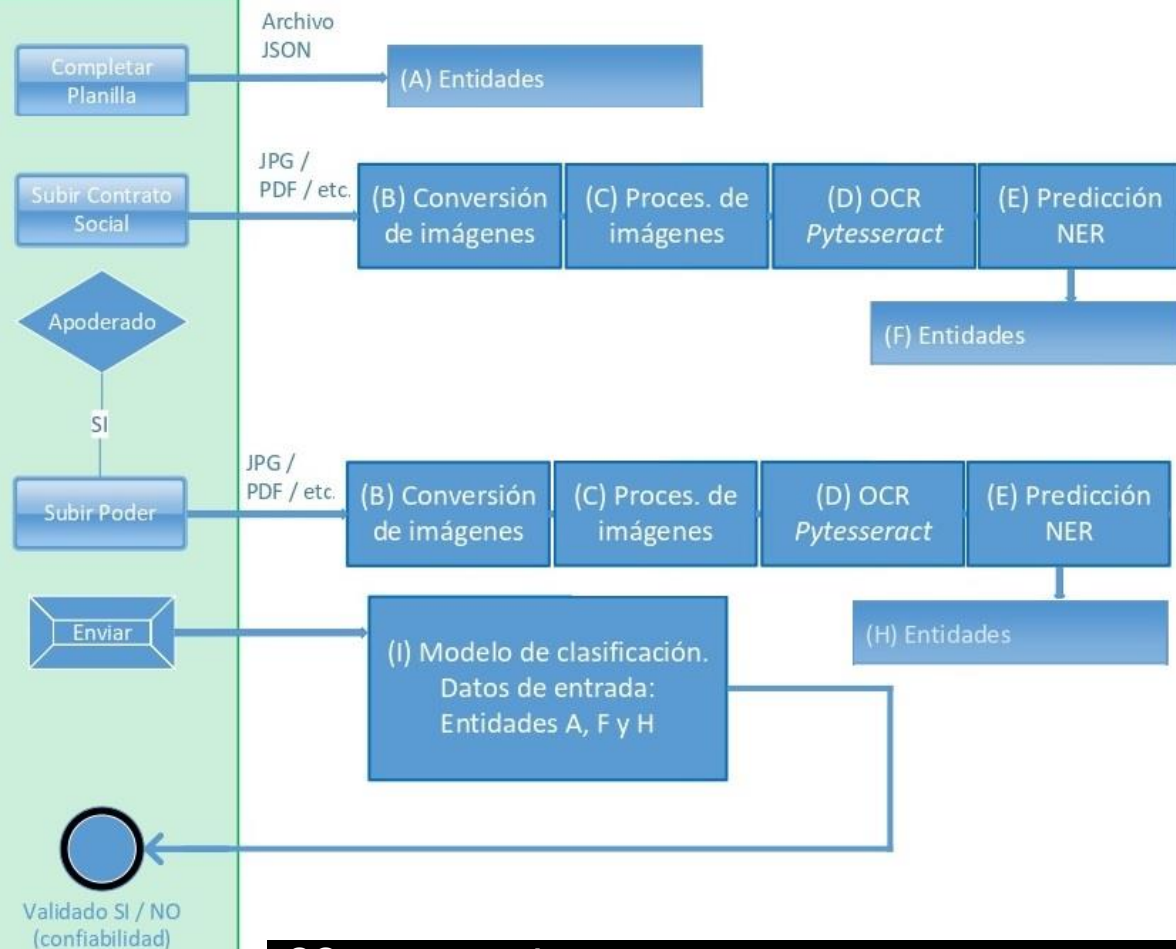
## VALIDACIÓN DE USUARIOS

Objetivos, desafíos y planteo general del sistema

# PROCESAMIENTO DE CONTRATOS



## INTERFAZ



## ESQUEMA GENERAL DEL SISTEMA

### (A) Entidades

A1: Razón Social

A2: Integrantes

A3: DNI/CUIT

A4: Nacionalidad

A5: Representante legal

A6: Fecha de registro

A7: Nro. de registro

# 03

## PROCESAMIENTO DE IMAGENES

Preprocesamiento, *Bounding Box* y OCR

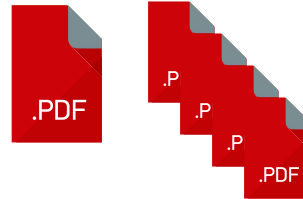


# PREPROCESAMIENTO

CONVERSIÓN  
DE FORMATO



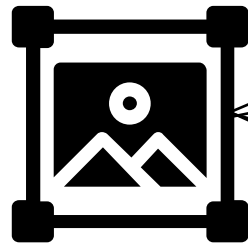
PAGINADO



CORRECCIÓN DE  
ÁNGULOS



# BOUNDING BOX



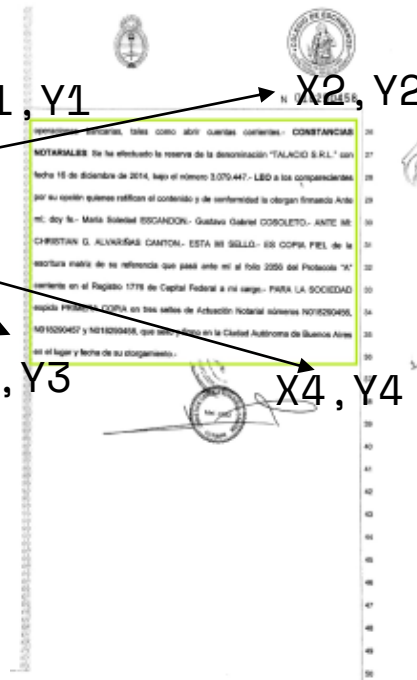
I.A.

X1, Y1

X2, Y2

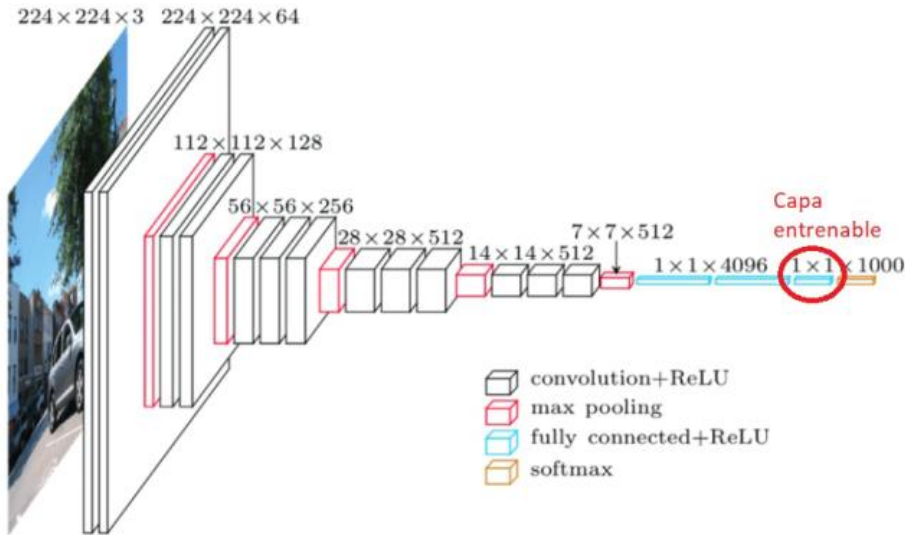
X3, Y3

X4, Y4

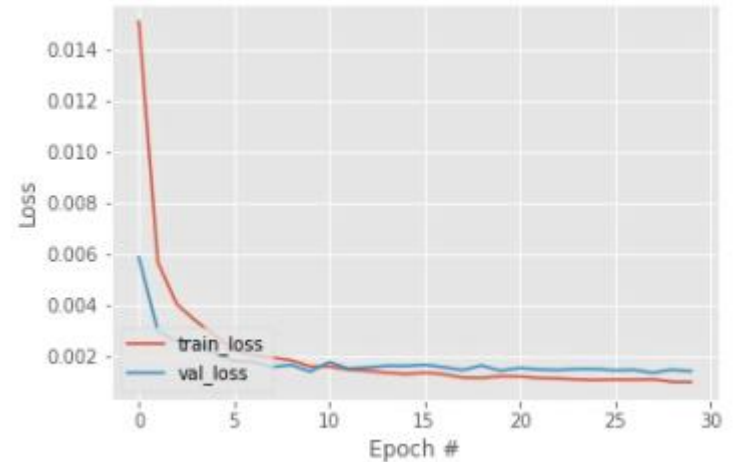


operaciones bancarias, tales como abrir cuentas corrientes.- CONSTANCIA  
NOTARIAL: Se ha efectuado la reserva de la denominación "TALACIO S.R.L." con  
fecha 16 de diciembre de 2014, bajo el número 3.079.447.- LBO a los comparecientes  
por su propia voluntad ratifican el contenido y de conformidad la otorgan firmados Ante  
mi, doy fe.- María Soledad ESCOBAR.- Gustavo Gabriel COSOLETO.- ANTE MI  
CHRISTIAN D. ALVAREZ CANTON.- ESTA MI SELLO.- ES COPIA FIEL de la  
escritura matriz de su referencia que pasó ante mí el folio 2550 del Protocolo "X"  
conforme en el Registro 1779 de Capital Federal a mi cargo.- PARA LA SOCIEDAD  
expido PRIMERA COPIA en tres volúmenes de Actas Notariales números N°18290455,  
N°18290457 y N°18290458, que están y están en la Ciudad Autónoma de Buenos Aires  
en el lugar y fecha de su otorgamiento.-

# BOUNDING BOX



MODELO VGG16 CON *FINETUNING*



ENTRENAMIENTO

# OCR TESSERACT

01		PREPROCESAMIENTO		Mejoras sobre la imagen que optimizan los pasos siguientes.
02		SEGMENTACIÓN		Identificación de regiones que contienen caracteres individuales.
03		EXTRACCIÓN		Obtención de las características de cada región.
04		RECONOCIMIENTO		Clasificación de cada caracter mediante etiquetado.
05		CORRECCIÓN		Postprocesamiento para corregir errores buscando coherencia.

# 04

## PROCESAMIENTO DE LENGUAJE NATURAL

NER (*Named Entity Recognition*)

# DEFINICIÓN DE ENTIDADES Y ETIQUETADO

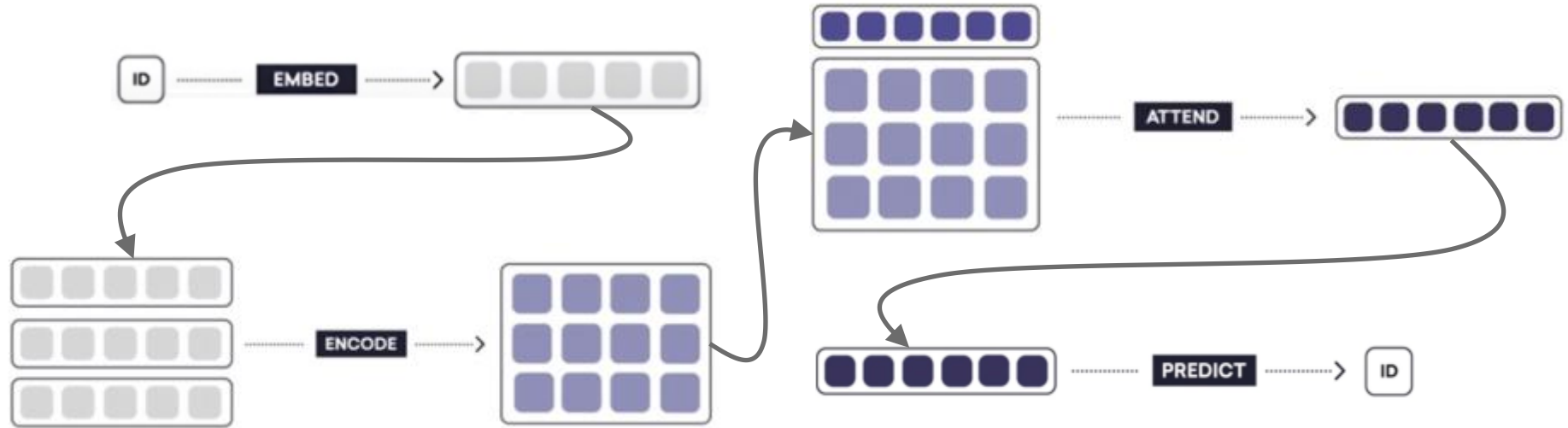
- Razón social.
- Integrantes de la sociedad.
- DNI/CUIT de los integrantes.
- Nacionalidad de los integrantes.
- Representante legal.
- Fecha de registro de la sociedad.
- Número de registro de la sociedad.

ENTIDADES

Orden FOM/2427/2012 **LEGAL** , de 29 de octubre **TIME** , por la  
empresas navieras para percibir las correspondientes bonificaciones al  
Autónomas de **Canarias Loc** y de las **Illes Balears Loc** .  
El **Real Decreto 1316/2001 LEGAL** , de 30 de noviembre **TIME**  
transporte aéreo y marítimo para los residentes en las Comunidades Au  
Ciudades de **Ceuta Loc** y **Melilla Loc** , establece en su dispos

ETIQUETADO CON PRODIGY

# PIPELINE DEL MODELO DE SPA.CY



A decorative header element consisting of two solid black squares, one on the left and one on the right, framing the central text.

# 05

## CLASIFICACIÓN BINARIA

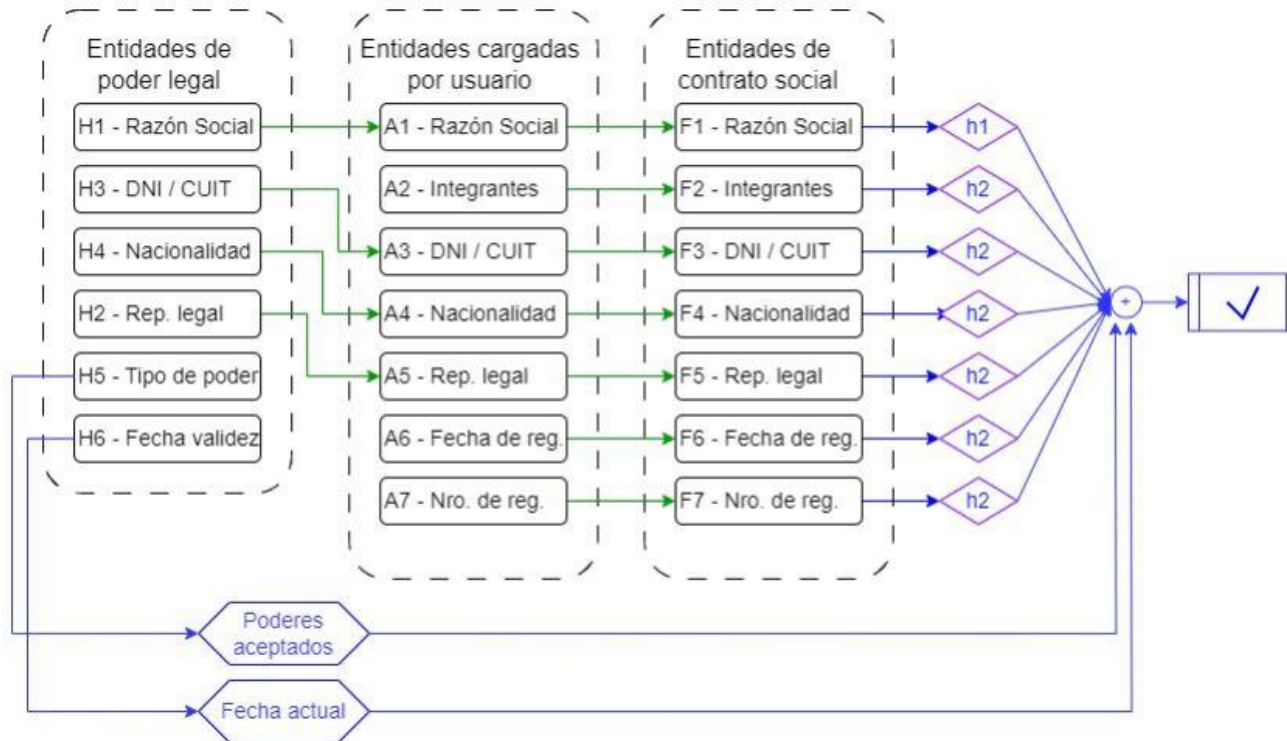
---

Algoritmo y diagrama de flujo



# CLASIFICADOR BINARIO

## DIAGRAMA DE FLUJO



# ALGORITMO DE CLASIFICACIÓN

## DISTANCIA LEVENSHTTEIN

Computa:

- Sustitución
- Inserción

Ejemplo:

1. Casa > cala (sustitución de "s" por "l")
  2. Cala > calla (inserción de "l")
  3. Calla > calle (sustitución de "a" por "e")
- Distancia entre casa y calle = 3

## IMPLEMENTACIÓN UTILIZADA

Python Weighted-levenshtein

## VARIANTE CON PESOS


Tiene en cuenta la similitud de caracteres.

Ejemplo:

`dist(0 , 0) < dist(0 , F)`

```
pip install weighted-levenshtein
```





06

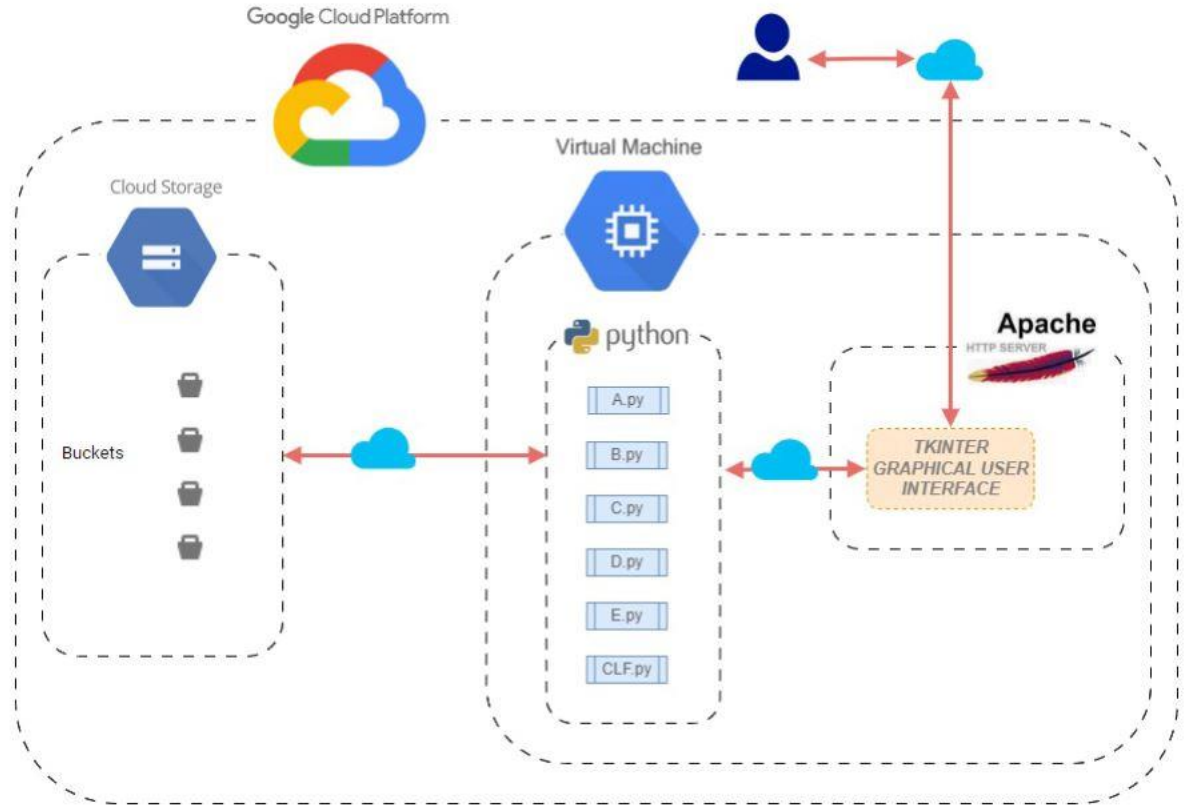
# IMPLEMENTACIÓN

---

Arquitectura, video y resultados



# ARQUITECTURA DE LA SOLUCIÓN

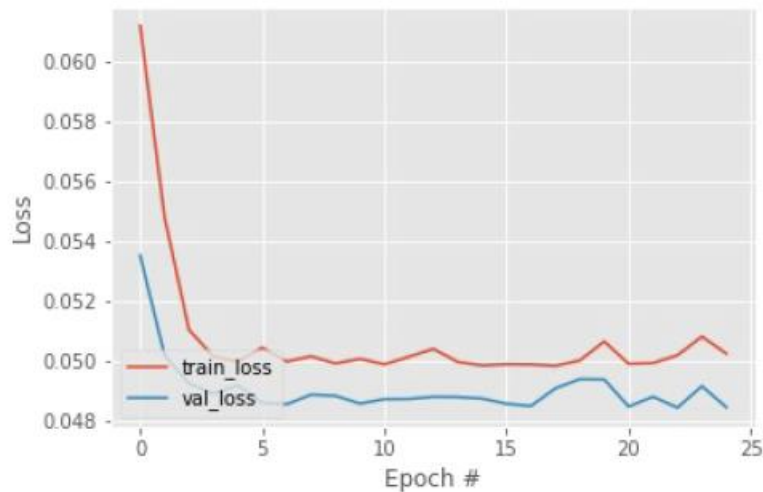


# VIDEO DE DEMOSTRACIÓN

---

<https://youtu.be/LPmeurBBOHQ>

# RESULTADOS DE MODELO VGG16



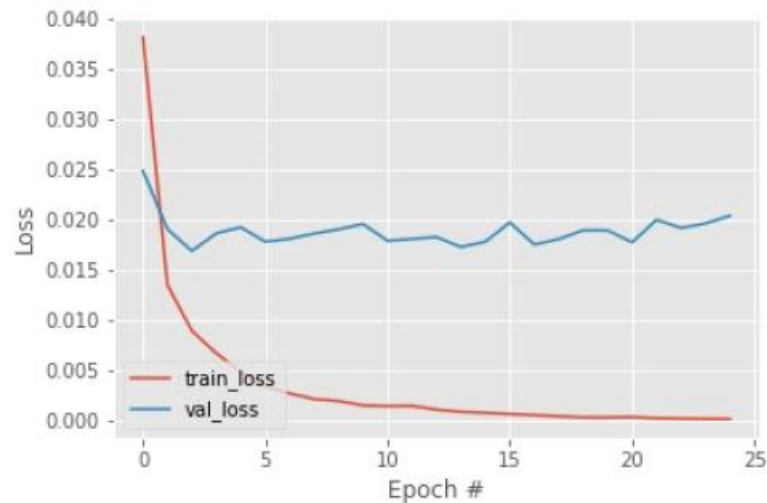
## Capas entrenables:

- Últimas tres convolucionales
- Última flatten (1,25.088)

LR: 0.0004

Épocas: 25

Lote: 32



## Capas entrenables:

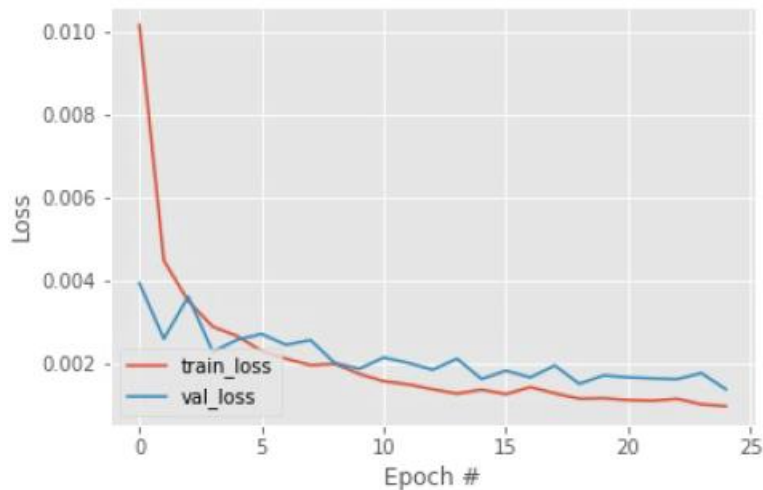
- Últimas tres convolucionales
- Última flatten (1,25.088)
- 4 capas densas adicionales

LR: 0.0004

Épocas: 25

Lote: 32

# RESULTADOS DE MODELO VGG16

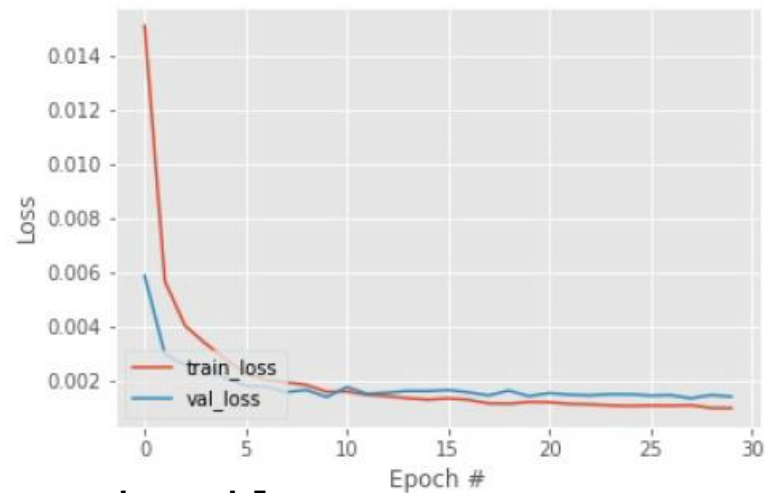


## Capas entrenables:

- Últimas tres convolucionales
- Última flatten (1,25.088)
- 4 capas densas adicionales
- Capa de dropout de 0.2

LR: 0.0004

Épocas: 25 Lote: 6



## Capas entrenables:

- Últimas tres convolucionales
- Última flatten (1,25.088)
- 4 capas densas adicionales
- Capa de dropout de 0.2

LR: 0.0004 variable "ReduceLR0Plateau"

Épocas: 25 Lote: 6

# RESULTADOS DE MODELO NLP

Training pipeline							
Components: ner							
Merging training and evaluation data for 1 components							
- [ner] Training: 82   Evaluation: 20 (20% split)							
Training: 67   Evaluation: 17							
Labels: ner (7)							
# Pipeline: ('tok2vec', 'ner')							
# Initial learn rate: 0.001							
E	#	LOSS TOK2VEC	LOSS NER	ENTS_F	ENTS_P	ENTS_R	SCORE
0	0	0.00	1448.63	0.23	0.12	1.62	0.00
2	200	15991.61	34462.76	53.61	73.86	42.07	0.54
5	400	2392.57	4632.48	56.16	74.73	44.98	0.56
8	600	22250.09	4087.05	49.22	78.17	35.92	0.49
11	800	998.98	2618.62	63.22	73.71	55.34	0.63
14	1000	1441.97	1690.21	62.12	74.89	53.07	0.62
17	1200	1339.32	1375.62	52.94	75.45	40.78	0.53
20	1400	889.96	1068.16	54.33	71.81	43.69	0.54
23	1600	3262.87	896.95	61.26	78.68	50.16	0.61
26	1800	3906.34	834.77	61.93	70.83	55.02	0.62
29	2000	856.94	678.83	60.69	81.87	48.22	0.61
32	2200	1087.92	680.12	57.20	74.87	46.28	0.57
35	2400	960.35	518.07	62.26	67.29	57.93	0.62







07

# EPÍLOGO

---

Conclusiones y trabajos futuros



# CONCLUSIONES

---



RESTRICCIONES EN EL ACCESO A DATOS



DATASET RESTRINGIDO



RESULTADO GLOBAL DE BAJA USABILIDAD



DISPONIBILIZACIÓN EXITOSA



BUENA PERFORMANCE DEL MODELO DE BOUNDING BOXES

# TRABAJO FUTURO

---



AMPLIACIÓN DEL DATASET



OPTIMIZACIÓN DEL RENDIMIENTO DE LOS MODELOS



EXPLORACIÓN DE TÉCNICAS AVANZADAS DE PROCESAMIENTO DE LENGUAJE NATURAL



EVALUACIÓN CONTINUA DE RESULTADOS

**GRACIAS !**

---

PREGUNTAS?