Modelos de Lenguaje Natural para Clasificación de Desagregaciones de Robo

Lenin G. Falconí

Escuela Politécnica Nacional

Quito, Noviembre 2024

QR-CODE Presentación

Temario

Contenido

Objetivo General

Desarrollar un modelo de aprendizaje automático para la clasificación del texto de la Noticia del Delito en las desagregaciones de Robo definidas por la Comisión Especial de Estadística de Seguridad, Justicia, Crimen y Transparencia(CEESJCT).

Formulación del Problema

Se propone entrenar un modelo de Machine Learning que permita *aproximar* una función paramétrica f_θ que a partir de un dato de entrada $\mathbf{x}_i \in \mathcal{X}$ devuelva la categoría k_j a la que pertenece dicho dato. Donde $k_j \in \mathcal{Y}$ y $\mathcal{Y} \in \mathbb{R}^{11}$

$$f_\theta : \mathcal{X} \longrightarrow \mathcal{Y} \tag{1}$$

Sobre el relato se aplica la Tokenización (i.e. *encoding*) del modelo pre-entrenado

$$\mathbf{x}_i = \Gamma(relato_i) \tag{2}$$

La tokenización genera la representación en *embeddings* del texto con dos tensores: 1) tokenización y 2) máscara de atención

Delitos Validados

Idx	Tipo de Robo	idx	Tipo de Robo
0	ROBO A INSTITUCIONES EDUCATIVAS	6	ROBO A UNIDADES ECONOMICAS
1	ROBO DE MOTOS	7	ROBO A DOMICILIO
2	ROBO EN INSTITUCIONES PUBLICAS	8	ROBO DE BIENES, ACCESORIOS Y AUTOPARTES DE VEHICULOS
3	ROBO DE CARROS	9	ROBO A EMBARCACIONES DE ESPACIOS ACUATICOS
4	ROBO A ESTABLECIMIENTOS DE COLECTIVOS U ORGANIZACIONES SOCIALES	10	OTROS ROBOS
5	ROBO A PERSONAS		

Cuadro: Delitos Validados

Contenido

Outline

Objetivos

1. Analizar la similitud documental de una muestra de relatos de policía $\mathbb{D}^{PN}$ wrt. al relato de fiscalía $\mathbb{D}^{FGE}$.
2. Utilizar al menos dos técnicas distintas en naturaleza: 1) word2vec, 2) Transformers
3. Calcular el coseno de la similitud entre los embeddings i.e.: $\cos(\Gamma(x_i^{PN}), \Gamma(x_i^{FGE})) = \frac{\Gamma(x_i^{PN}) \cdot \Gamma(x_i^{FGE})}{\|\Gamma(x_i^{PN})\| \|\Gamma(x_i^{FGE})\|}$, donde $x_i^{PN} \in \mathbb{D}^{PN}$ y $x_i^{FGE} \in \mathbb{D}^{FGE}$

Metodología I

1. Calcular la muestra n a utilizar.
2. Tomar n elementos aleatorios desde el conjunto de datos de relatos N, si la cantidad de palabras es al menos 50: $E(w) \geq 50$.
3. Definir los modelos o técnicas para el análisis del texto: 1) word2vec, 2) Transformers
4. Obtener los *embeddings* para cada par de relatos de policía y fiscalía i.e. $\Gamma(x_i^{PN}), \Gamma(x_i^{FGE})$ de acuerdo a cada técnica seleccionar.
5. Obtener el coseno de similitud
6. Comparar resultados estadísticamente
7. Realizar comparación de las predicciones del Modelo entrenado en *delitos seguimiento* sobre relatos de policía y fiscalía y comparar resultados de clasificación en función de las métricas de clasificación.

Cálculo de la Muestra

$$n = \frac{z^2 p(1-p)}{\epsilon^2 N + z^2 p(1-p)} N$$

Donde:

$N = 785\,513$

$p = 0{,}5$

$z = 1{,}65$

$\epsilon = 1{,}06\,\%$

$n = 6\,012 \approx 6\,000$

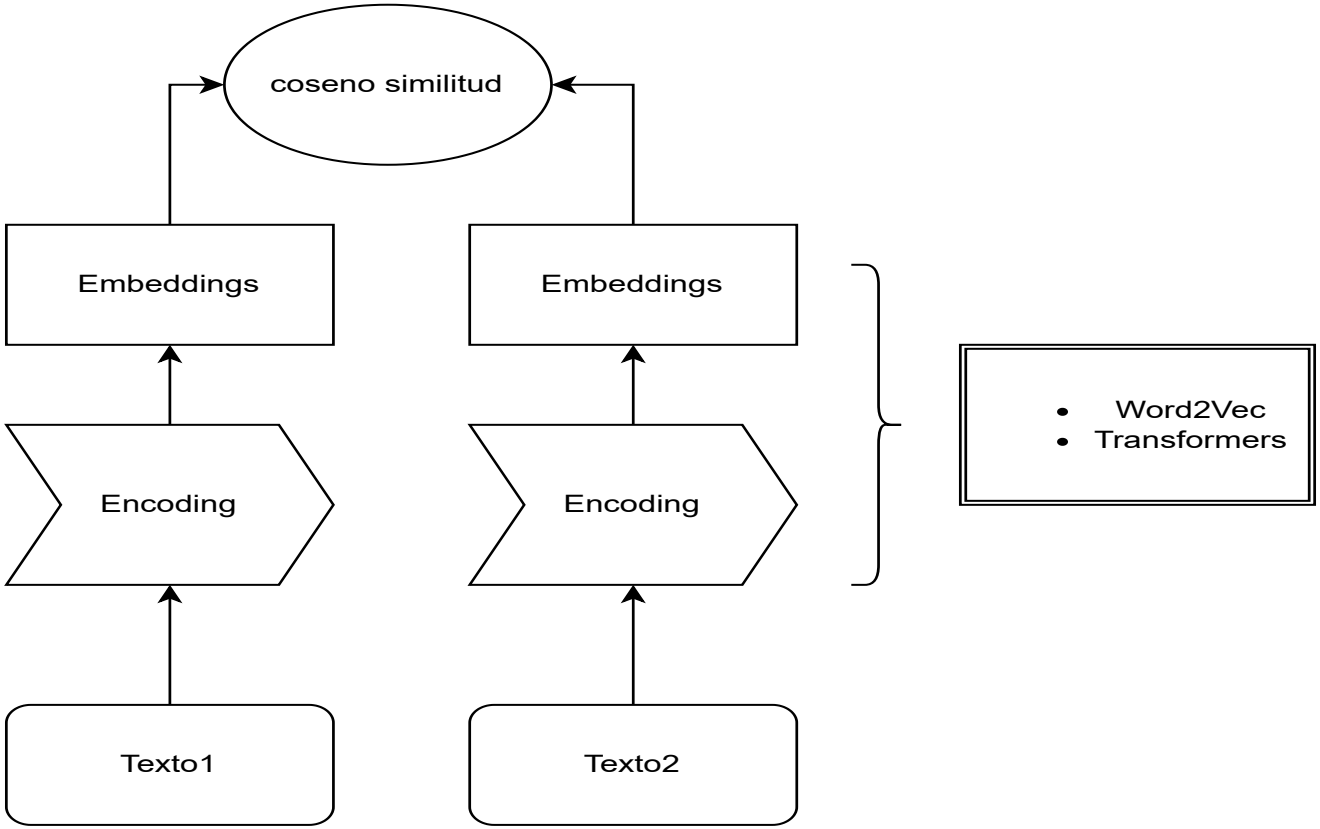
Enlace Calculadora Muestra

Se requiere de una muestra de 6 000 registros para presentar resultados con una confianza del 90 % de un total de 785 513.

Similitud Documental

Cálculo

text embedding: son representaciones numéricas vectoriales de información (e.g. texto, imagen, audio,e tc.) en un espacio dimensional menor. Captura el significado semántico de la información.



Similitud Documental

Resultados

	word2vec	bert-transformer
count	6 000	6 000
mean	0,957 418	0,987 899
std	0,073 233	0,023 865
min	0,310 206	0,780 015
25 %	0,961 602	0,990 691
50 %	0,975 714	0,995 396
75 %	0,984 985	0,997 255
max	1	1

Predicción Delitos Seguimiento en Relatos

Formulación

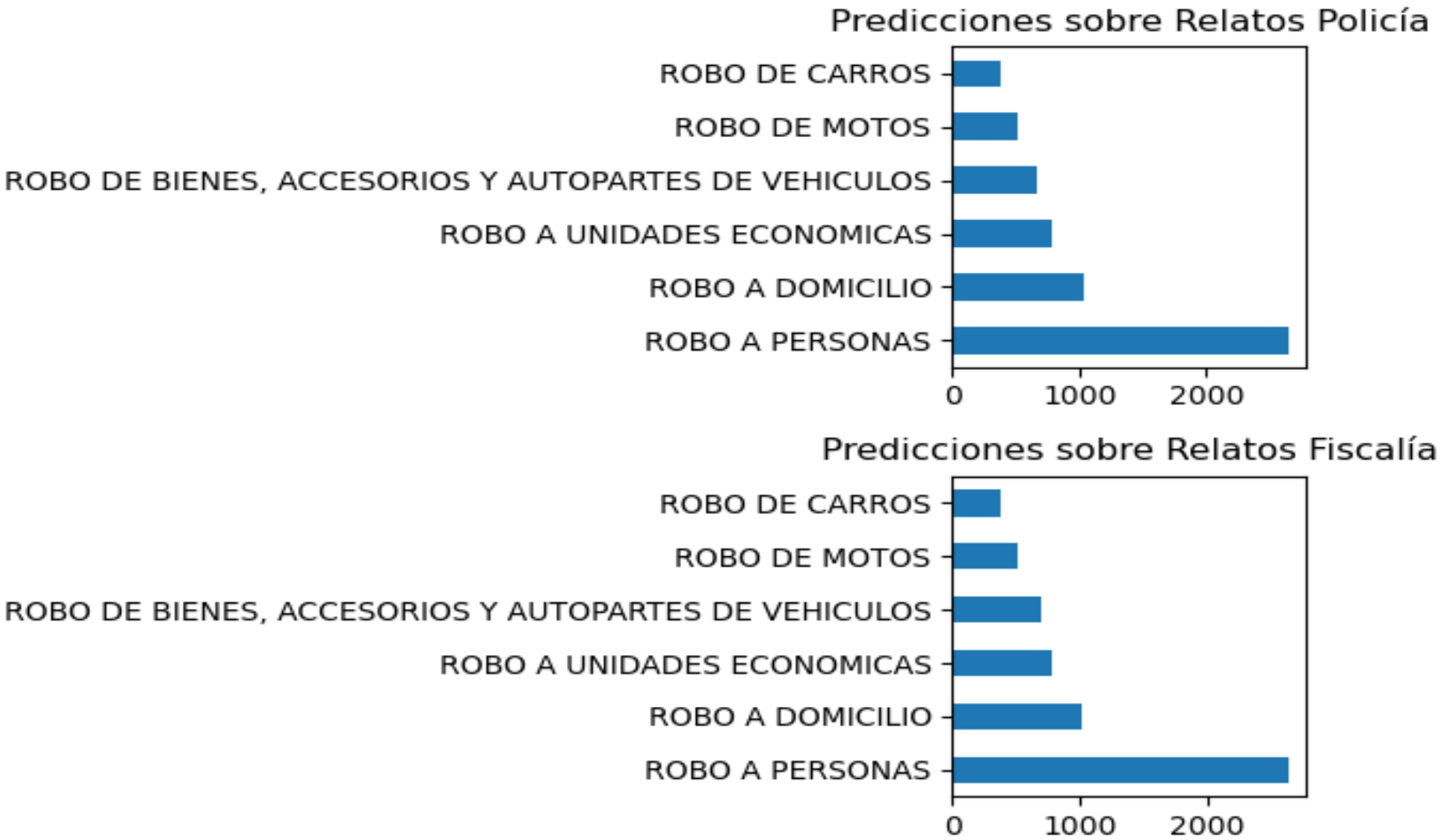
Se realiza la predicción del modelo entrenado de *delitos seguimiento* sobre los textos de los relatos de policía y fiscalía i.e.

$$y^{PN} = f_\theta(\mathcal{X}^{PN})$$

$$y^{FGE} = f_\theta(\mathcal{X}^{FGE})$$

Predicción Delitos Seguimiento en Relatos

Resultados



Predicción Delitos Seguimiento en Relatos Fiscalía

Resultados

Categoría	Precision	Recall	f1-score	soporte
ROBO A DOMICILIO	0.80	0.92	0.86	564
ROBO A PERSONAS	0.90	0.95	0.93	1458
ROBO A UNIDADES ECONOMICAS	0.61	0.88	0.72	245
ROBO DE BIENES, ACCESORIOS Y AUTOPARTES DE VEHICULOS	0.85	0.95	0.89	493
ROBO DE CARROS	0.93	0.93	0.93	312
ROBO DE MOTOS	0.98	0.95	0.96	415
accuracy			0.86	3807
macro avg	0.72	0.80	0.76	3807
weighted avg	0.80	0.86	0.83	3807

Cuadro: Reporte de Clasificación de predicciones sobre relato Fiscalía

Predicción Delitos Seguimiento en Relatos Policía

Resultados

Categoría	Precision	Recall	f1-score	soporte
ROBO A DOMICILIO	0.80	0.93	0.86	564
ROBO A PERSONAS	0.90	0.95	0.93	1458
ROBO A UNIDADES ECONOMICAS	0.58	0.88	0.70	245
ROBO DE BIENES, ACCESORIOS Y AUTOPARTES DE VEHICULOS	0.86	0.93	0.89	493
ROBO DE CARROS	0.89	0.90	0.90	312
ROBO DE MOTOS	0.97	0.93	0.95	415
accuracy			0.85	3807
macro avg	0.72	0.79	0.75	3807
weighted avg	0.79	0.85	0.82	3807

Cuadro: Reporte de Clasificación de predicciones sobre relato Policía

Conclusiones I

- Para fines computacionales los relatos de policía y de fiscalía son similares, con una media de 0,97 del coseno de similitud.
- El 75 % de los datos de la muestra supera el 95 % de similitud.
- La diferencia global entre las predicciones del modelo de *delitos seguimiento* f_θ sobre los relatos de policía y fiscalía es de 0,16 %.
- Con respecto a las métricas de evaluación del clasificador, se observa un rendimiento general de 0,86 para el relato de Fiscalía y 0,85 para el relato de Policía.
- La inferencia y capacidad de generalización del modelo no está supeditada al origen del relato.
- De lo anterior, se confirma que la manera de mejorar al modelo no está en entrenar con u origen particular del texto del relato (i.e. policía o fiscalía), sino en entrenar con la mayor cantidad de datos posibles; lo que implica un mayor consumo de hardware

Outline

Metodología I

1. Lectura de registros de CEESJCTde 2014 a 2022 donde Tipo Penal es Robo con las columnas NDD, Tipo Delito PJ, delitos seguimiento y delitos validados i.e. $\mathbb{D}^{A}_{735045\times 4}$
2. Obtención de relatos fiscalía y policía $\forall NDD_i \in \mathbb{D}^{A}_{735045\times 4}$ desde la base de relatos $\mathbb{D}^{FGE\cup PN\cup CE}$: $\mathbb{D}^{B}_{725079\times 8}$
3. Se excluyen tipos penales diferentes de robo i.e. $\mathbb{D}^{B'}_{723435\times 8} \subset \mathbb{D}^{B}_{725079\times 8}$
4. $\mathbb{D}^{robos}_{723435\times 11} = \mathbb{D}^{A}_{735045\times 4} \bowtie_{\theta} \mathbb{D}^{B'}_{723435\times 8}$
5. Formateo de relato: texto a minúsculas y retirar caracteres que no sean letras o números.
6. Integrar la columna $relato_{pn}$ y $relato_{fge}$ en una sola columna de relato y conservando las columnas de NDD, Presunto Delito, Relato, delitos seguimiento y delitos validados i.e. $\mathbb{D}^{robos}_{1446870\times 5} = \mathbb{D}^{robos_{pn}}_{723435\times 5} \cup \mathbb{D}^{robos_{fge}}_{723435\times 5}$

Metodología II

7. Ordenar ascendente según valor de NDD a $\mathbb{D}^{robos}_{1446870\times5}$
8. Analizar la estadística de la cantidad de palabras de los relatos de policía y fiscalía. Se conserva los relatos tales que: $50 \leq len(relato_i) \leq 400$. Pues, límite superior (*upper fence*) del q_3 del relato de policía es 389,5 i.e. $\mathbb{D}^{robos}_{(1140728\times6)}$
9. Corregir etiquetas de delitos seguimiento: se sustituye las letras tildadas por sus equivalentes sin tilde económicas $\rightarrow$ economicas
10. Se obtiene, entonces, $\mathbb{D}^{robos}_{(1109335\times6)}$ con las categorías definidas en Tabla 1.
11. Las filas del dataset se orden aleatoriamente
12. Se salva los datos en sql:

```
from src.utils import save_df_in_sql
save_df_in_sql(dataf=dataset_out,name_table='
    dataset_RobosDesagregation06122023')
```

Metodología III

Notación:

$\bowtie_\theta$: Inner Join donde $\theta : \mathbb{D}^A.NDD = \mathbb{D}^B.NDD$

Categorías del Dataset

delitos_seguimiento	delitos_validados	Total
ROBO A DOMICILIO	ROBO A DOMICILIO	172264
ROBO A PERSONAS	ROBO A PERSONAS	421497
ROBO A UNIDADES ECONOMICAS	ROBO A UNIDADES ECONOMICAS	74088
ROBO DE BIENES, ACCESORIOS Y AUTOPARTES DE VEHICULOS	ROBO DE BIENES, ACCESORIOS Y AUTOPARTES DE VEHICULOS	154546
ROBO DE CARROS	ROBO DE CARROS	90038
ROBO DE MOTOS	ROBO DE MOTOS	119128
SIN INFORMACION	OTROS ROBOS	43468
	ROBO A EMBARCACIONES DE ESPACIOS ACUATICOS	9407
	ROBO A ESTABLECIMIENTOS DE COLECTIVOS U ORGANIZACIONES SOCIALES	3087
	ROBO A INSTITUCIONES EDUCATIVAS	17252
	ROBO EN INSTITUCIONES PUBLICAS	4560
Total		1 109 335

Cuadro: Categorías de Delitos Validados y Seguimiento en el Dataset

Organización de datos de Entrenamiento, Validación y Testo I

1. Los datos se organizan de la siguiente manera:
   - Entrenamiento: 807 468 $\approx$ 72 %
   - Validación: 201 867 $\approx$ 18 %
   - Test: 100 000 $\approx$ 10 %
2. El dataset se separa en dos subconjuntos dependiendo si las etiquetas corresponden a delitos seguimiento o delitos validados
3. Se conserva únicamente el relato y la etiqueta (i.e. delitos_validados) que se renombra como *labels*.
4. Los datos se guardan en la base de datos de *Machine Learning*:

Organización de datos de Entrenamiento, Validación y Testo II

```
from src.utils import save_df_in_sql
save_df_in_sql(
     dataf=train_delitos_validados_huggingface,
    database="machinelearning",
    name_table="train_delitos_validados_hf",
)
save_df_in_sql(
    dataf=valid_delitos_validados_huggingface,
    database="machinelearning",
    name_table="valid_delitos_validados_hf",
)
save_df_in_sql(
    dataf=test_delitos_validados_huggingface,
    database="machinelearning",
    name_table="test_delitos_validados_hf",
)
```

Outline

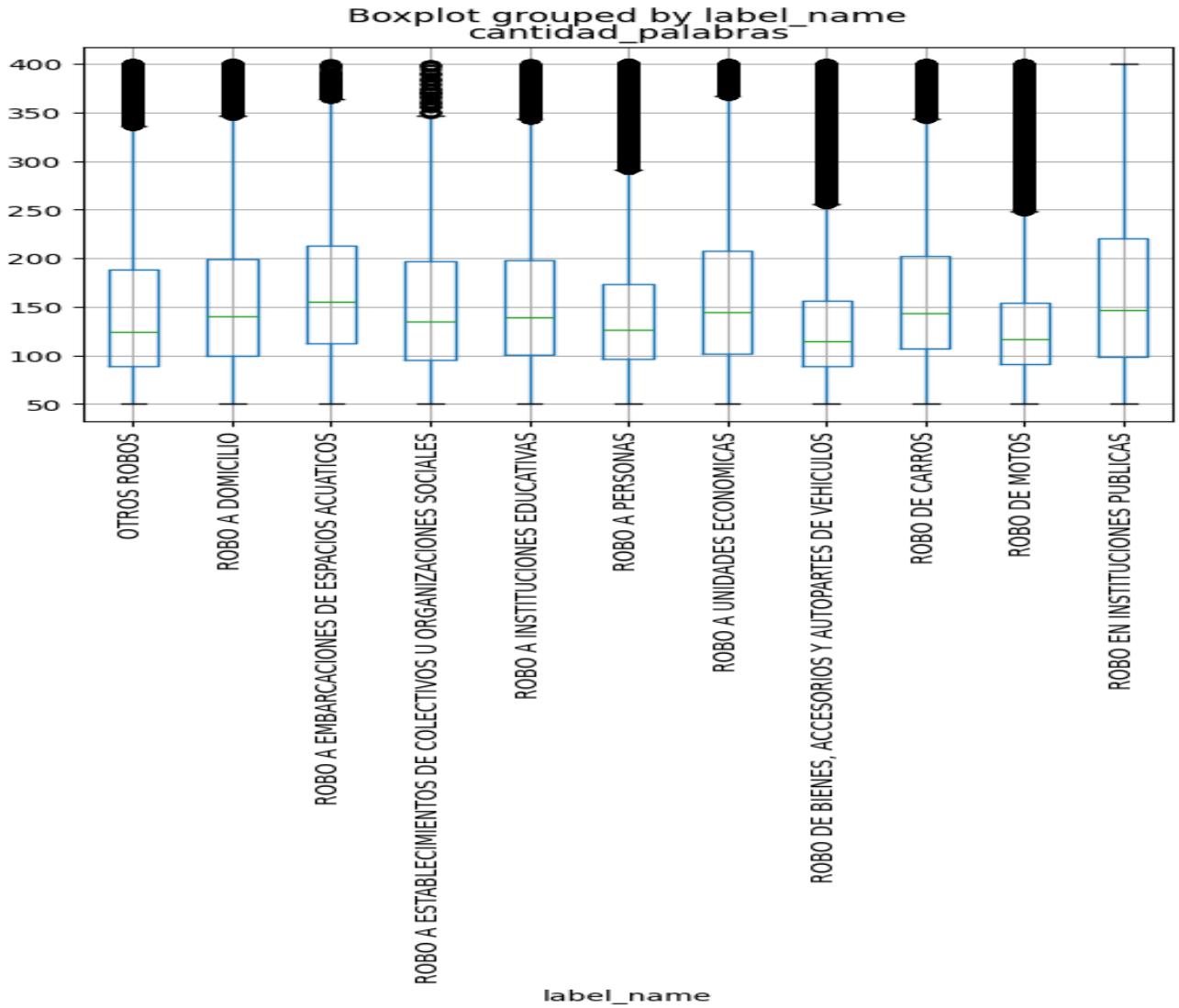
Entrenamiento del Modelo I

Metodología

1. Carga de los datos de entrenamiento $\mathcal{X}^{Train}_{807468\times 2}$, validación, $\mathcal{X}^{Valid}_{201867\times 2}$ y testeo $\mathcal{X}^{Test}_{100000\times 2}$
2. Tokenización de los relatos con una secuencia máxima de 400 y empleando el modelo distilbert-base-multilingual-cased
3. Habilitación de Tensor Processing Unit
4. Configuración de earlystopping con *patience* $= 10$, con monitoreo del *accuracy* de validación y retorno de los mejores pesos.
5. Configuración de Hiperparámetros:
   - $Batch\_size = NumeroReplicas \times 16 = 8 \times 16 = 128$
   - El número de épocas se configura en 12
   - El optimizador a usar es Adam con *learning rate* de 3×10^{-6}

Entrenamiento del Modelo

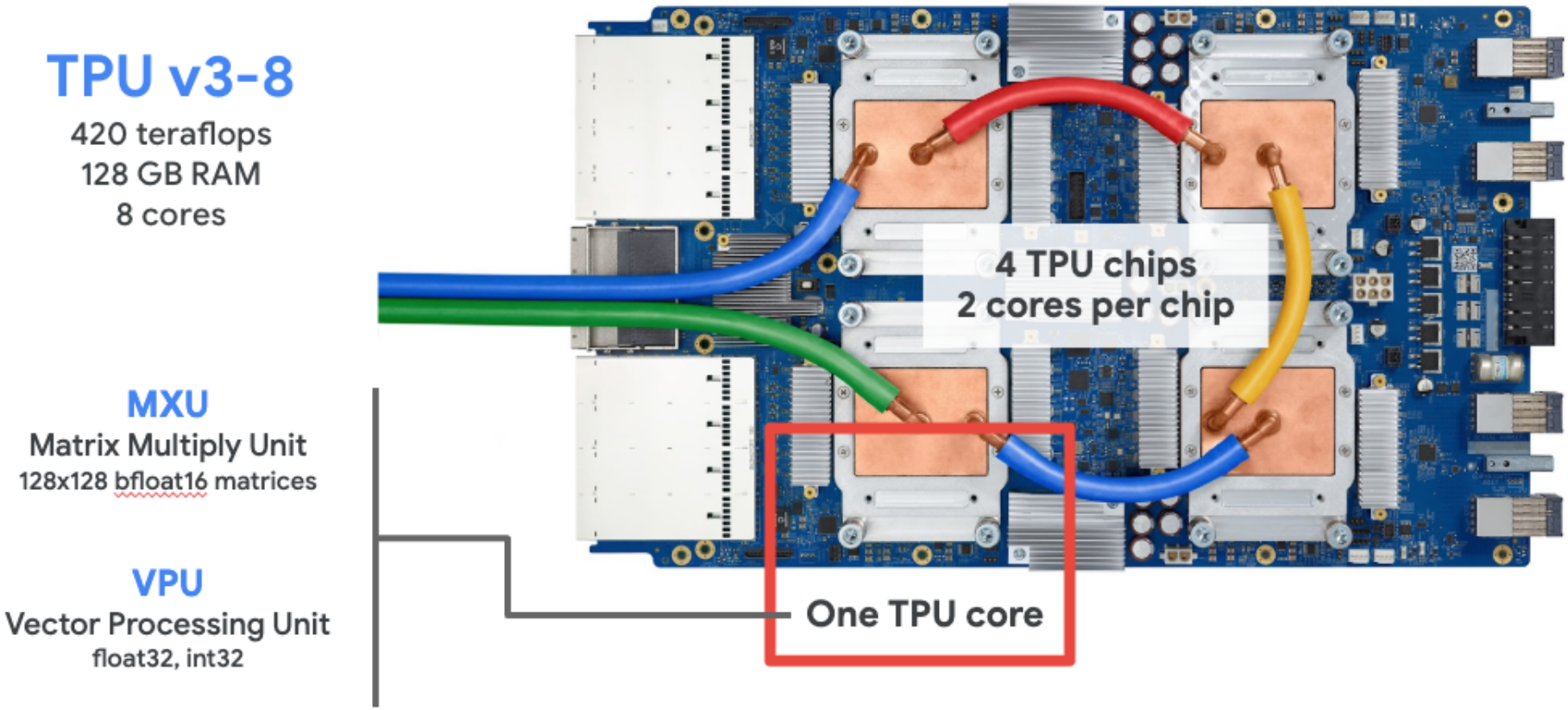
Características del relato de $\mathcal{X}^{Train}$



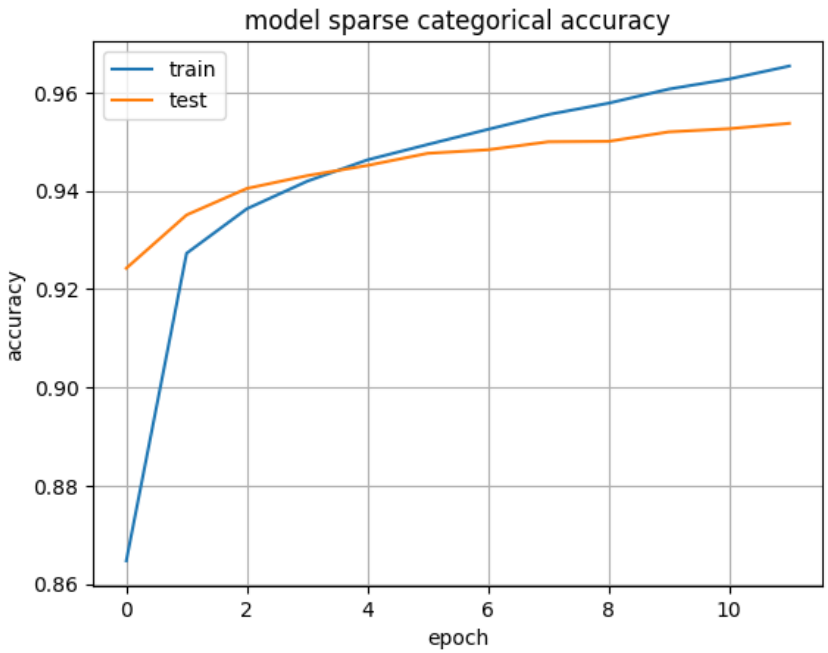
Entrenamiento del Modelo

TPU

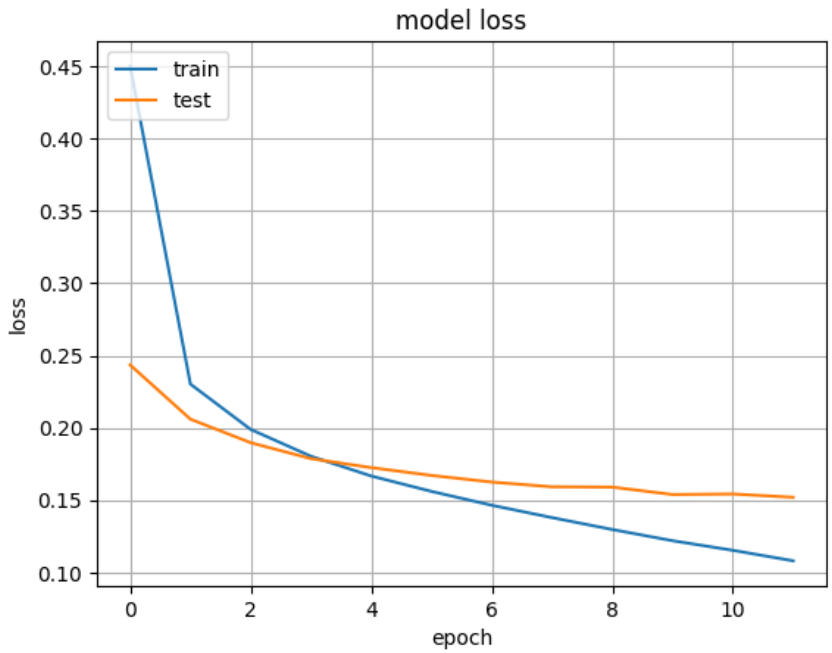
TPU: Acelerador de hardware para *Deep Learning*



Resultados de Entrenamiento



Accuracy por Época

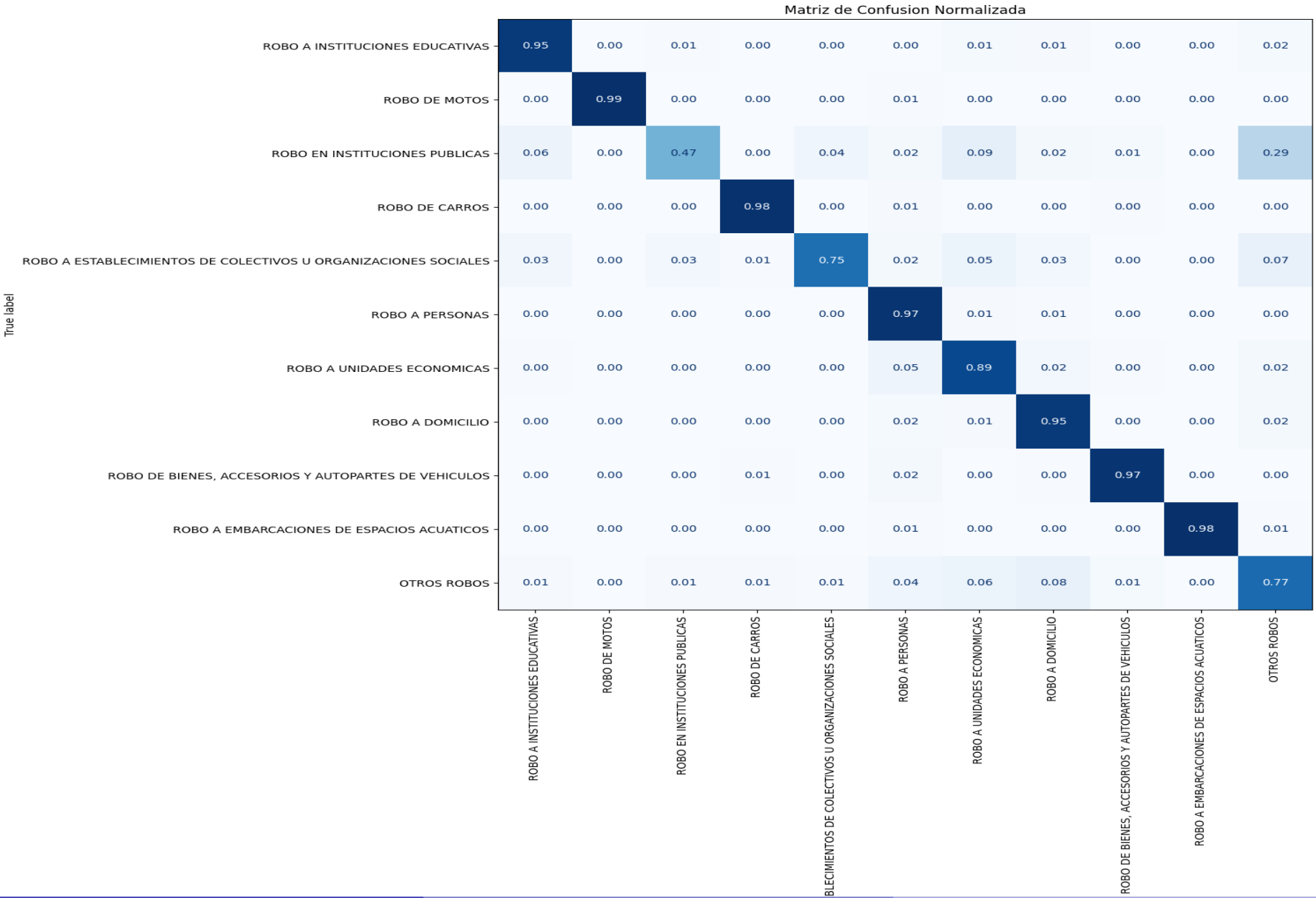


Loss por Época

Accuracy y Loss de Validación

Outline

Matriz de Confusión Normalizada



Reporte de Clasificación

	precision	recall	f1-score	support
ROBO A INSTITUCIONES EDUCATIVAS	0.925995	0.946599	0.936184	1573
ROBO DE MOTOS	0.985211	0.988317	0.986762	10785
ROBO EN INSTITUCIONES PUBLICAS	0.724528	0.474074	0.573134	405
ROBO DE CARROS	0.967669	0.980198	0.973893	7878
ROBO A ESTABLECIMIENTOS DE COLECTIVOS U ORGANIZACIONES SOCIALES	0.750000	0.752688	0.751342	279
ROBO A PERSONAS	0.969847	0.972298	0.971071	37976
ROBO A UNIDADES ECONOMICAS	0.878871	0.893877	0.886311	6794
ROBO A DOMICILIO	0.944779	0.952335	0.948542	15504
ROBO DE BIENES, ACCESORIOS Y AUTOPARTES DE VEHICULOS	0.977752	0.968340	0.973023	14024
ROBO A EMBARCACIONES DE ESPACIOS ACUATICOS	0.965197	0.983452	0.974239	846
OTROS ROBOS	0.817738	0.766006	0.791027	3936
accuracy			0.954610	100000
macro avg	0.900690	0.879835	0.887775	100000
weighted avg	0.954050	0.954610	0.954175	100000

Conclusiones I

- El nuevo modelo de *delitos validados* mejora el desempeño de precisión alcanzando un valor de 0,954 6. En consequencia, tiene un mejor rendimiento que el modelo de *delitos seguimiento*
- A nivel computacional se ha probado con dos técnicas diferentes y los textos del parte policial y del relato siaf son similares.
- El desempeño mejorado del modelo de *delitos validados*, que predice 11 categorías, se debe a, como se había señalado, que se pudo entrenar con una mayor cantidad de ejemplos. Gracias al uso de TPU.
- Entre las diferentes métricas de *precision*, *recall* y *f1-score* se observa que existe un rendimiento adecuado del clasificador en las distintas categorías.

Conclusiones II

- La categoría con el desempeño más bajo corresponde a "ROBO EN INSTITUCIONES PUBLICAS". Sin embargo, como puede observarse en Tabla 4, es la segunda menor en frecuencia con 4 560 registros luego de "ROBO A ESTABLECIMIENTOS DE COLECTIVOS U ORGANIZACIONES SOCIALES" que tiene 3 087.
- De la Matriz de Confusión puede observarse que existe una confusión aproximadamente del 29 % entre "ROBO EN INSTITUCIONES PUBLICAS" y "OTROS ROBOS". Sin embargo, esta puede ser evidencia de que se deba revisar manualmente la categoría de "OTROS ROBOS" o en su defecto eliminar esta categoría e integrarla en *OTROS ROBOS*.