Universidad de Jaén

Clasificación Automática del Tipo de Delitos de Odio en Denuncias Policiales

Autor:
Juan Bautista Muñoz Ruiz

Tutor:
Prof. D.ª Salud María Jiménez Zafra
Prof. D.ª María Teresa Martín Valdivia

ujaen.es

INTRODUCCIÓN

- Según el “Informe sobre la evolución de los delitos de Odio” presentado en 2021 por el Ministerio del Interior:

  - Se **incrementaron** un **24%** el número de delitos de **racismo.**
  - **Aumentaron** en un **30%** los delitos hacia el **género** y preferencias sexuales.

Universidad de Jaén

INTRODUCCIÓN

- Como consecuencia de esta **tendencia al alza** y de la **digitalización acelerada** tras la pandemia.

- Es de interés **aplicar conocimientos y tecnologías** computacionales en este ámbito.

- Desarrollando un **Sistema Clasificador de Delitos de Odio.**

Universidad de Jaén

MOTIVACIÓN

- **Seguir aprendiendo** en campos como la IA, PLN, Machine Learning y Ciencia de Datos.

- **Automatizar una tarea** tan ardua como la clasificación de denuncias.

- Generar un producto de **valor y de gran utilidad.**

Universidad de Jaén

OBJETIVOS

- Realizar un **estudio** de:
  - Los diferentes modelos de clasificación.
  - Técnicas de Procesamiento del Lenguaje Natural.
  - Aprendizaje automático.

- **Diseñar y desarrollar** un sistema clasificador automático de delitos de Odio.

Universidad de Jaén

OBJETIVOS

- Diseñar e implementar una **aplicación** de visualización de resultados **intuitiva y sencilla**.

- Redactar la **memoria** del proyecto con toda la información.

- Redacción del **manual de instalación y usuario.**

Universidad de Jaén

PROPÓSITO

- Crear un **sistema clasificador** multifuncional:
  - **Entrenamiento** de modelos.
  - **Procesado** de denuncias policiales.
  - **Validación** de modelos.

- Crear una **plataforma web** para visualización de resultados:
  - Clasificación de denuncias en **formato escrito.**
  - Clasificación de denuncias en **formato .docx.**

Universidad de Jaén

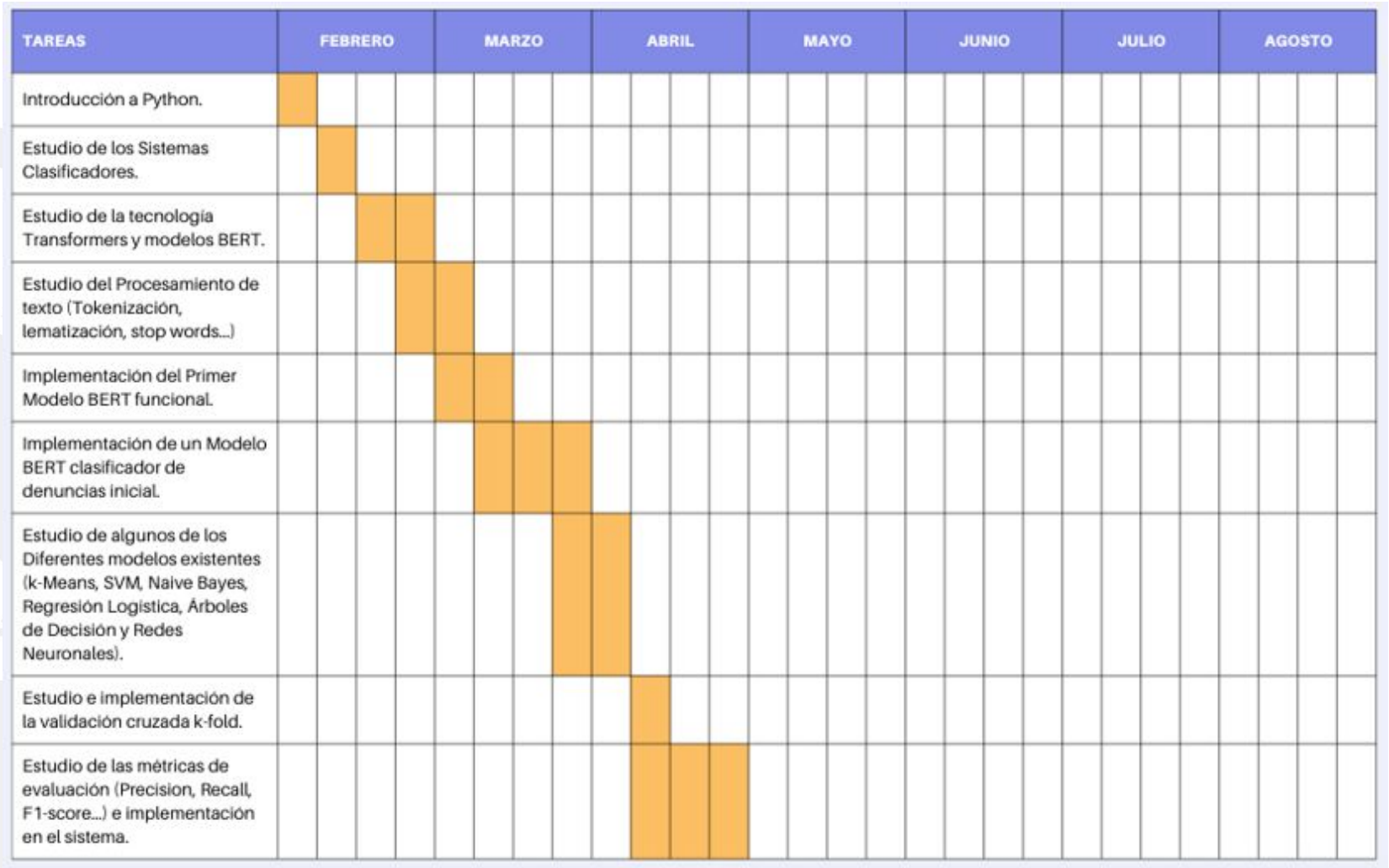
PLANIFICACIÓN

- **Metodología ágil**:
  - Creación de incrementos funcionales de valor.
  - Revisión y mejora continua de secciones ya implementadas.

- Solapamiento con el cursado de la asignatura **Procesamiento del Lenguaje Natural**.

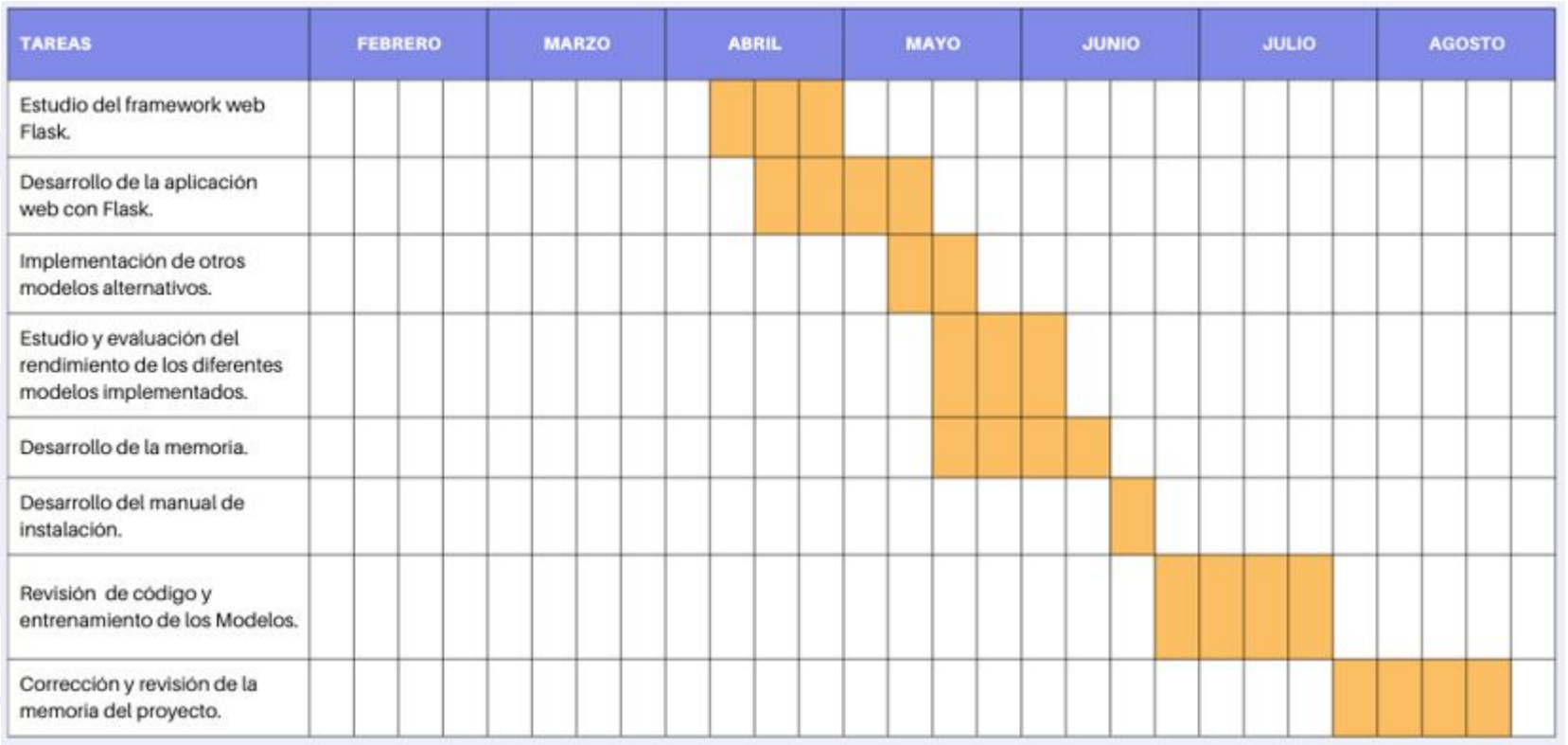
Universidad de Jaén

PLANIFICACIÓN



Universidad de Jaén

PLANIFICACIÓN



TAREAS	FEBRERO	MARZO	ABRIL	MAYO	JUNIO	JULIO	AGOSTO
Estudio del framework web Flask.							
Desarrollo de la aplicación web con Flask.							
Implementación de otros modelos alternativos.							
Estudio y evaluación del rendimiento de los diferentes modelos implementados.							
Desarrollo de la memoria.							
Desarrollo del manual de instalación.							
Revisión de código y entrenamiento de los Modelos.							
Corrección y revisión de la memoria del proyecto.							

Universidad de Jaén

COSTES

Presupuesto	Coste
Costes en software	249 €
Costes en hardware	1.345,61 €
Costes en recursos humanos	17.766 €
Costes no previstos	1.776 €
Total	**21.136,61 €**

ESTUDIOS REALIZADOS

- Corpus:

NOODIO Contiene: 80 archivos

anonimizado_0500_Caso_C0740394_obj_78677408.PDF.docx
anonimizado_0500_Caso_C0740398_obj_78695538.PDF.docx
anonimizado_0500_Caso_C0740401_obj_79053671.PDF.docx
anonimizado_0500_Caso_C0740402_obj_78717651.PDF.docx

ANONIMIZADO_0500_Caso_C0740394_obj_78677408.PDF

Atestado: AT<<<<<< Instructor: NUM<< Secretario: NUM<< Atestado nº: AT<<<<<< Dependencia: DEPE<<<<<<<<<<

DILIGENCIA INICIAL: Se extiende, a las 02 horas 03 minutos del día 2019<<<<<<<<<<<<<<, para hacer constar, que se recibe Atestado número XXXXX de 17- 05-19 de la Cª de Madrid-USERA, en la que ANON<<<<<<<<<<<<<<<<<<<<<<<<<<<< con DNI NUM<<<<<<, NACIDA EL XXXX ANON< hija de Imael y ANON<<<< con domicilio en XXXXXXXXXXXXX y teléfono NUM<<<<<8 ha interpuesto denuncia por MALOS TRATOS hacia su marido, ANON<<<<<<<<<<<<<<<<<<<<<<<<<<<<<con DNI NUM<<<<<<, nacido en ANON< el 01-04-53, hijo de ANON<<<<y ANON<<<<< con domicilio en XXXXXXXXXXX de Madrid y teléfono NUM<<<<<1. --Que por estos motivos, el Jefe de la ODAC procede a comisionar al indicativo XXX, compuesto por los agentes con carnets profesionales números XXXXX y XXXX, al domicilio del denunciado para su localización y detención, siendo esta gestión POSITIVA.-. CONSTE Y CERTIFICO.

ODIO Contiene: 110 archivos,

P02023AX2AA_2016006223-AnonimizadoV2.docx
P02023AX2AA_2016006650-AnonimizadoV2.docx
P02023AX2AA_2016008466-AnonimizadoV2.docx
P02023AX2AA_2016008578-AnonimizadoV2.docx

ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P02023AX2AA_2016006223.pdf

AT< Instructor: INS< Secretario: AT< Dependencia: DEPE< OF.DENUNCIAS

En LOC<, siendo las TEMP< TEMP< del TEMP<, ante el PER< y Secretario arriba mencionados. -- COMPARECE: En calidad de DENUNCIANTE, quien mediante DNI nº NUM< y CAR< quien acredita ser PER<, país de nacionalidad LOC<, varón, nacido en LOC<, el TEMP<, hijo de PER< y PER< , con domicilio en LOC<, teléfono NUM<, tlfno./móvil NUM<, y: -- MANIFIESTA: Que denuncia los hechos, que se detallan a continuación, ocurridos entre las TEMP<, del TEMP< y las TEMP<, del TEMP<, en Centro médico, LOC<. -- Que ha sido informado de la obligación legal que tiene de decir la verdad (Art.433 de L.E.Cr.) y de la posible responsabilidad penal en la que puede incurrir en caso de acusar o imputar falsamente a una persona una infracción penal o con temerario desprecio hacia la verdad (Art. 456 de Código Penal), simular ser responsable o víctima de una infracción penal

Universidad de Jaén

ESTUDIOS REALIZADOS

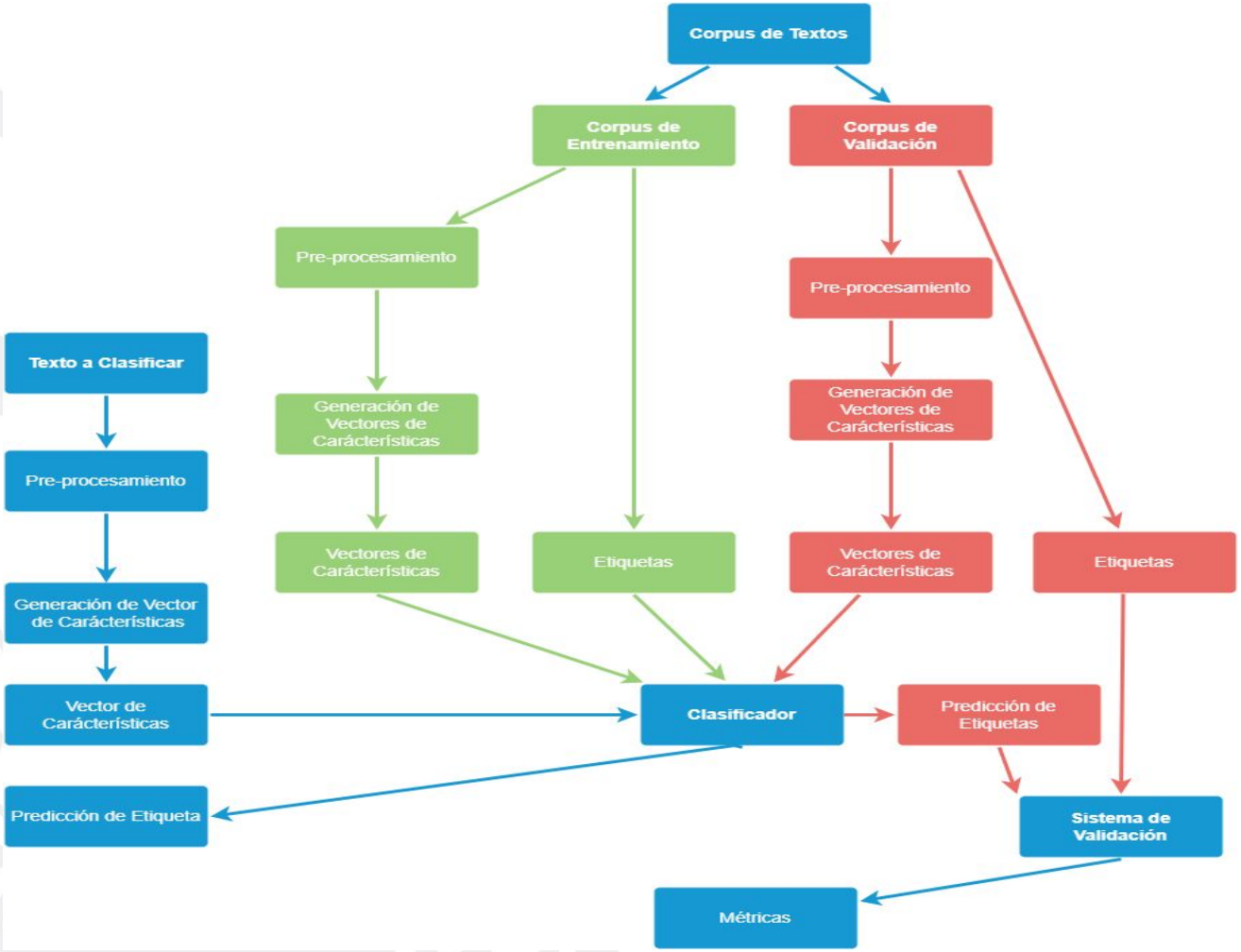
- Tipos de Sistemas clasificadores:
  - Según su objetivo: **Final** o No Final.
  - Según su orientación: **Procesamiento textual** o Comunicación Humano-Máquina.
  - Según el tipo de etiquetado: **Binario**, Multiclase o Multietiqueta.

- Estructura del sistema clasificador:

Universidad de Jaén

ESTUDIOS REALIZADOS

- Estructura del sistema clasificador:



Universidad de Jaén

ESTUDIOS REALIZADOS

- Técnicas de PLN:
  - **Tokenización.**
  - **Lematización** y Stemming
  - Vectores de características:
    - Basados en frecuencia: Bag of words, **TF-IDF** y Hash Vectorizer.
    - Basados en contexto: Word2vec.
- Tipos de modelos a usar: Basados en Reglas, **Machine Learning** o Híbridos.

Universidad de Jaén

ESTUDIOS REALIZADOS

- Tipos de modelos Machine Learning:
  - No supervisados:
    - K-Means.
  - **Supervisados**
    - NO BERT:
      - **SVM.**
      - **Naive Bayes.**
      - K-NN.
      - Regresión Lineal.
    - Redes Neuronales -> Transformers -> **Modelos BERT**:
      - BERT/**DISTIL-BERT** (**Uncased** o Cased).
      - **BERTIN.**
      - **BETO.**
      - MARIA.
      - MULTILINGUAL BERT.

Universidad de Jaén

ESTUDIOS REALIZADOS

- Validación Cruzada:
  - Leave-One.
  - **K-Fold:**

Sea un conjunto de archivos: [1,2,3,4,5,6,7,8,9] donde k=3:
Subconjuntos: {1,3,5} {7,8,9} {2,4,6}
Iteraciones:

- Iteración 1: Entrenamiento: {1,3,5} {7,8,9} Validación: {2,4,6}
- Iteración 2: Entrenamiento: {1,3,5} {2,4,6} Validación: {7,8,9}
- Iteración 3: Entrenamiento: {2,4,6} {7,8,9} Validación: {1,3,5}

Universidad de Jaén

ESTUDIOS REALIZADOS

- Métricas:

	Etiqueta	Predicha
Etiqueta Real	**A**	**B**
A	VP	FN
B	FP	VN

$$Precisión = \frac{VerdaderosPositivos}{VerdaderosPositivos \ + \ FalsosPositivos}$$

$$Recall = \frac{VerdaderosPositivos}{VerdaderosPositivos \ + \ FalsosNegativos}$$

$$F1\_Score = 2 \cdot \frac{Precisión \ \cdot \ Recall}{Precisión \ + \ Recall}$$

$$Acurracy = \frac{VerdaderoPositivos \ + \ VerdaderosoNegativos}{VerdaderoPositivos + VerdaderosoNegativos + FalsosPositivos \ + \ FalsosNegativos}$$

Universidad de Jaén

ESTUDIOS REALIZADOS

- Métricas:

	Etiqueta	Predicha
Etiqueta Real	**A**	**B**
A	VP	FN
B	FP	VN

$$Macro\ avg\ Métrica = \frac{Métrica\ etiqueta\ 1 + \ldots\ldots + Métrica\ etiqueta\ n}{n}$$

$$Weighted\ avg\ Métrica = \frac{Métrica\ etiqueta\ 1 \cdot Peso\ etiqueta\ 1 + \ldots\ldots + Métrica\ Etiqueta\ n \cdot Peso\ etiqueta\ n}{n}$$

Universidad de Jaén

DISEÑO

- 7 Requisitos Funcionales.

Requisitos Funcionales
RF1: El sistema debe clasificar una denuncia policial.
RF2: El sistema debe estar entrenado con el corpus de atestados policiales.
RF3: El sistema debe disponer de la capacidad de leer y procesar las denuncias policiales del corpus.
RF4: El sistema debe contar con la capacidad de entrenar diferentes modelos de clasificación.
RF5: El sistema debe tener un validador que genere unas métricas para medir el rendimiento de los diferentes modelos.
RF6: El sistema debe contar con una forma de selección de entre todos los modelos entrenados disponibles.
RF7: El sistema debe contar con una página web para la visualización de los resultados.

- 12 Requisitos No funcionales.

Requisitos No Funcionales
RNF1: La interfaz del sistema debe cumplir los principios de usabilidad.
RNF2: El tiempo de respuesta para la clasificación de una denuncia debe ser inferior a 4 segundos.
RNF3: El tiempo de respuesta para mostrar un mensaje de error en la interfaz web debe ser instantáneo.
RNF4: El espacio ocupado por parte de los modelos entrenados debe ser inferior a 10 GB
RNF5: El sistema debe contar con una interfaz amigable e intuitiva.
RNF6: El tiempo de entrenamiento de un modelo debe ser inferior a 5 horas.
RNF7: El tiempo de respuesta para cambiar de una vista de la web a otra debe ser menor a 1 segundo.

Requisitos No Funcionales
RNF8: La vista web debe hacer uso de metáforas como botones, ventanas de escritura, etc…
RNF9: La página web no tendrá tolerancia a fallo durante la subida o escritura de un archivo en formato incorrecto, notificándose así al usuario.
RNF10: El sistema clasificador no tendrá tolerancia al fallo durante un entrenamiento o una validación, notificandolo así por pantalla y deteniendo la ejecución
RNF11: Las vistas web han de ser web responsive para facilitar su uso y accesibilidad.
RNF12: El tiempo de lectura y procesado de denuncias debe ser menor que el de entrenamiento.

- 8 Historias de Usuario.

Como	**Usuario**
Quiero	Poder acceder a una página web
Para	Visualizar los resultados de la clasificación

Como	**Usuario**
Quiero	Poder escribir un texto en la página web
Para	Clasificar una denuncia policial en formato escrito

Como	**Analista Programador**
Quiero	Entrenar modelos distintos a BERT
Para	Realizar comparaciones

Universidad de Jaén

DISEÑO

- Diagramas:

Diagrama de clases del clasificador

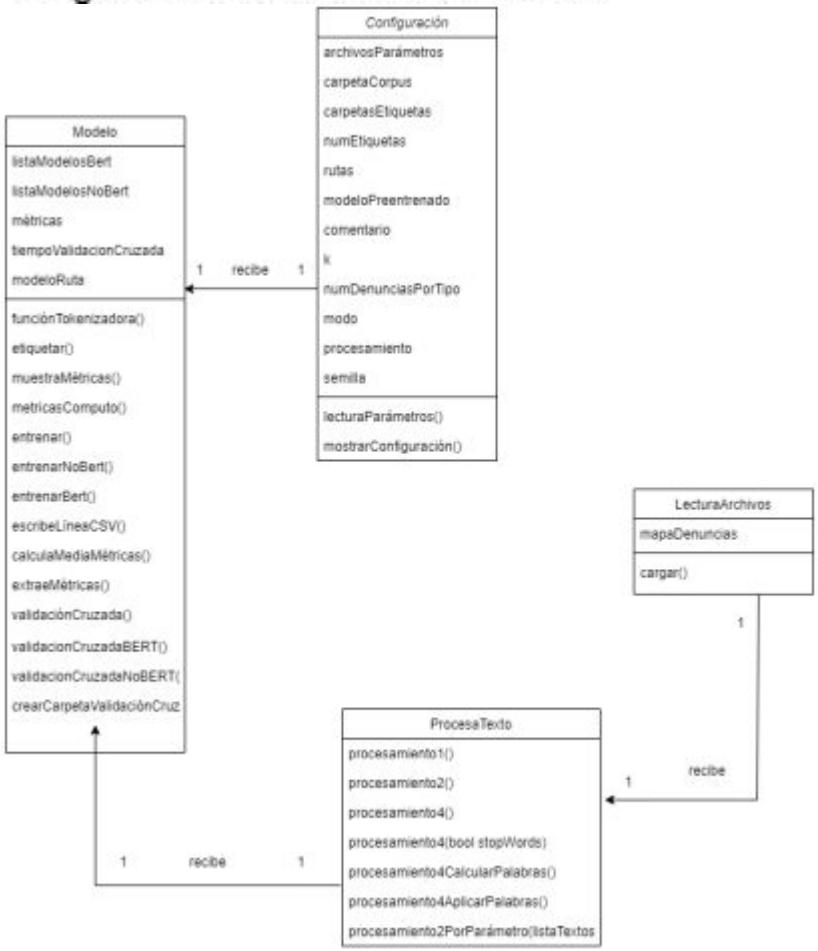


Diagrama de la Arquitectura MVC

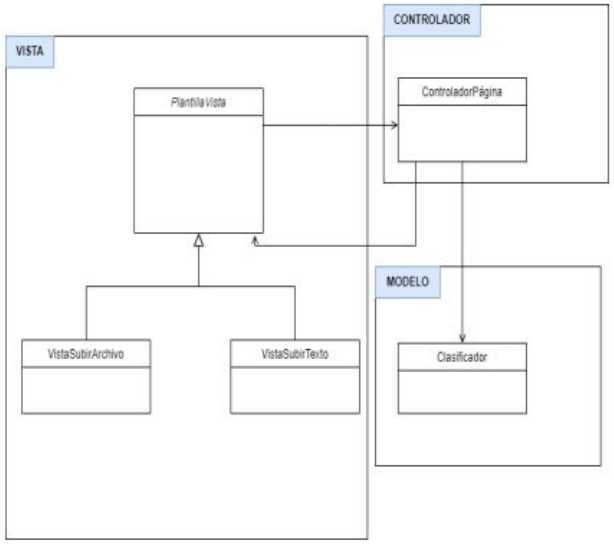
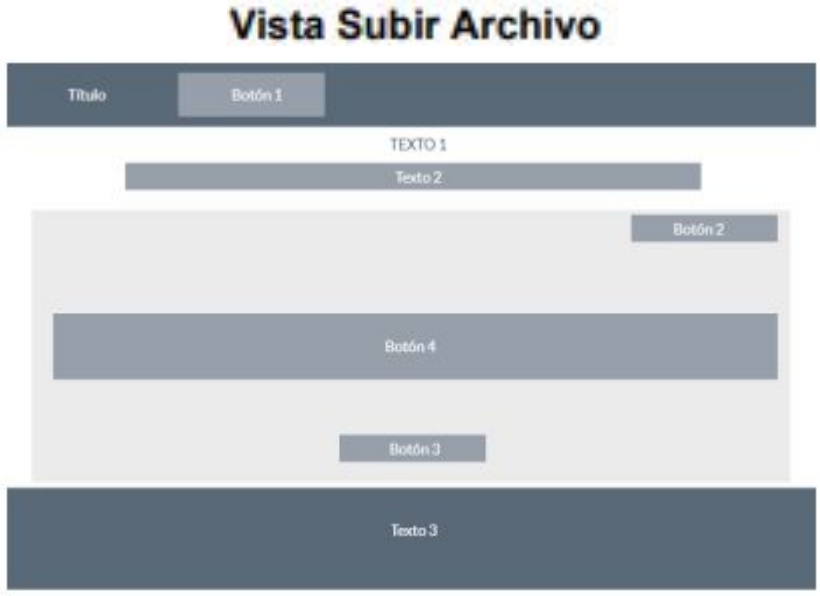


Diagrama de Clases de la Web

DISEÑO

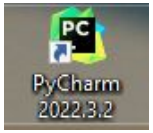
- Wireframes de la web:

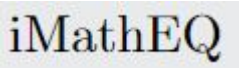


Universidad de Jaén

DESARROLLO

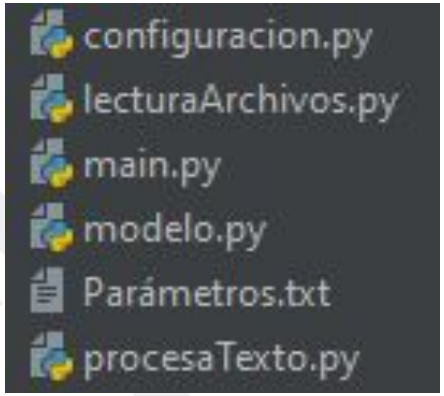
- Herramientas:

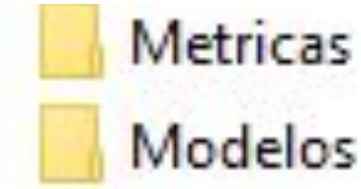




DESARROLLO

- Resultados Obtenidos (Modelo Clasificador):





Parámetros.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
carpetaCorpus=.\Corpus
modeloPreentrenado=distilbert-base-uncased
comentario=Vídeo
k=3
numDenunciasPorTipo=-1
modo=Validacion
procesamiento=2
semilla=13
```

modo=Validacion

modo=Entrenamiento

Procesamiento 1:
- **procesamiento=1**

Procesamiento 2:
- **procesamiento=2**

Procesamiento 3:
- **procesamiento=3**

Procesamiento 4a:
- **procesamiento=4**

Procesamiento 4b:
- **procesamiento=5**

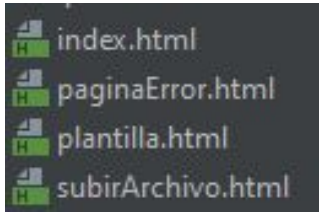
numDenunciasPorTipo=75

numDenunciasPorTipo=-1

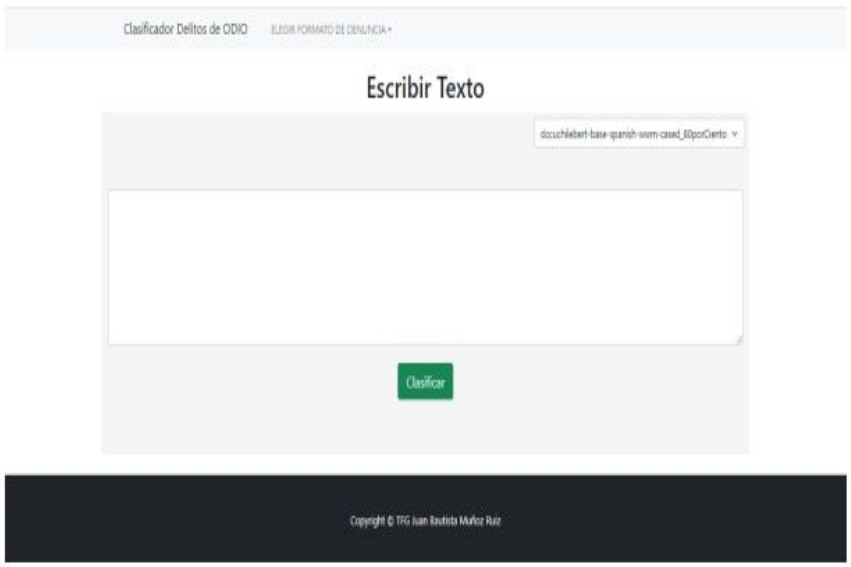
Universidad de Jaén

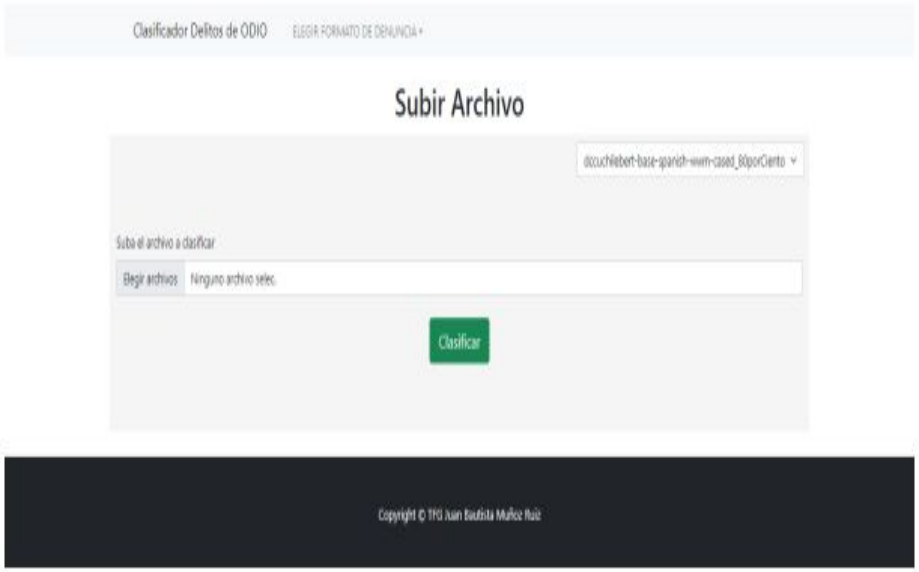
DESARROLLO

- Resultados Obtenidos (Web):



http://127.0.0.1:5000/.





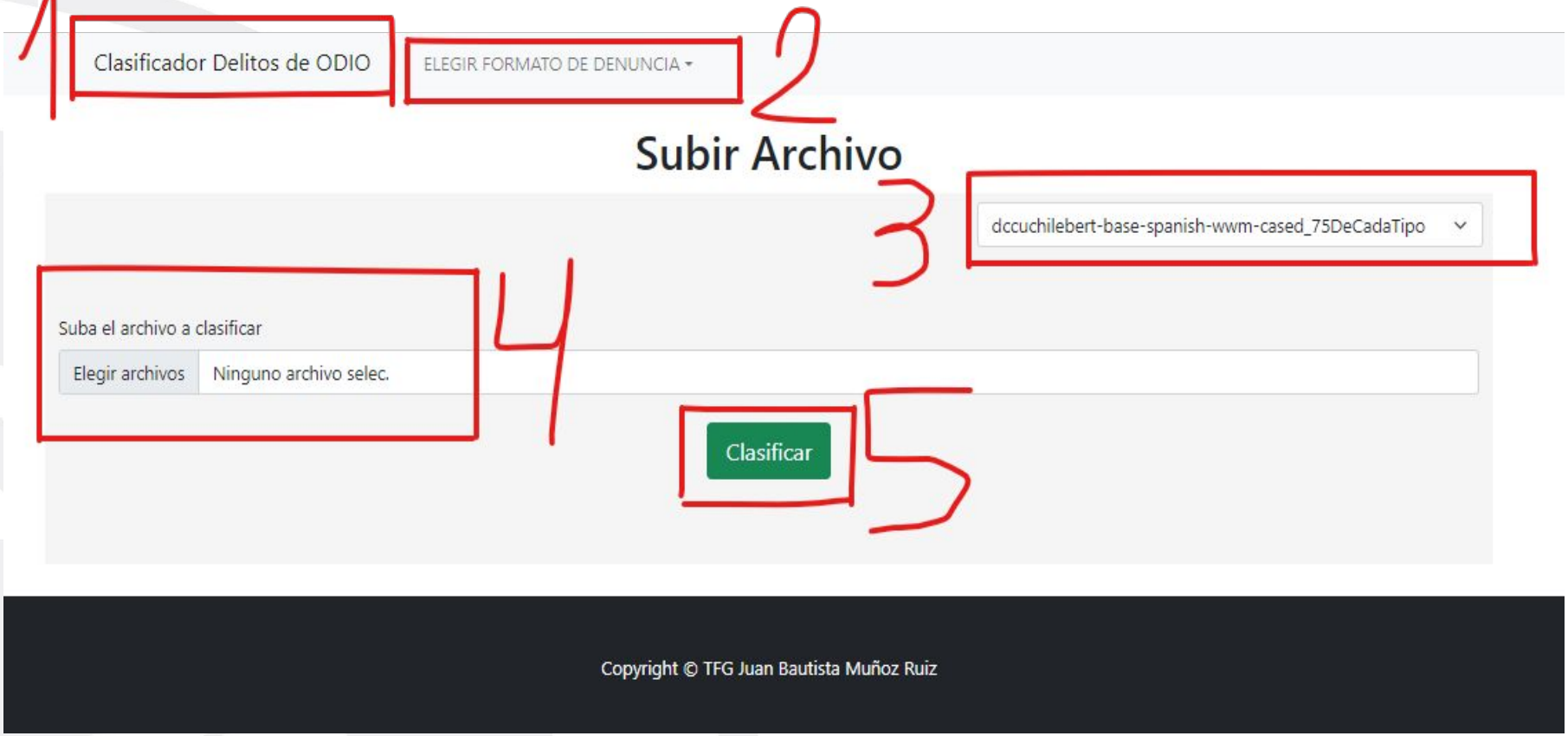
DESARROLLO

- Resultados Obtenidos (Web):

Universidad de Jaén

DESARROLLO

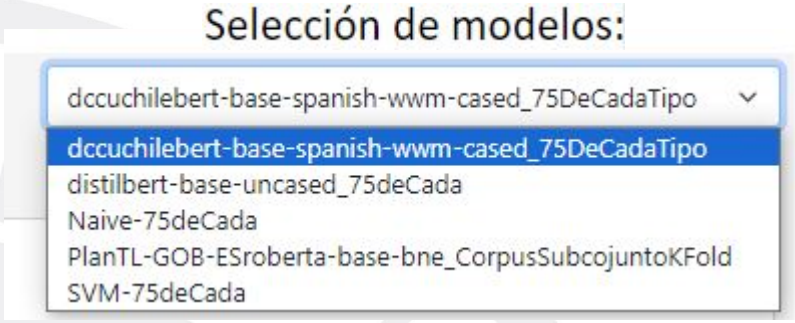
- Resultados Obtenidos (Web):

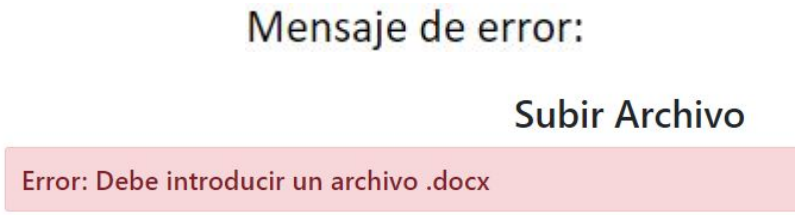


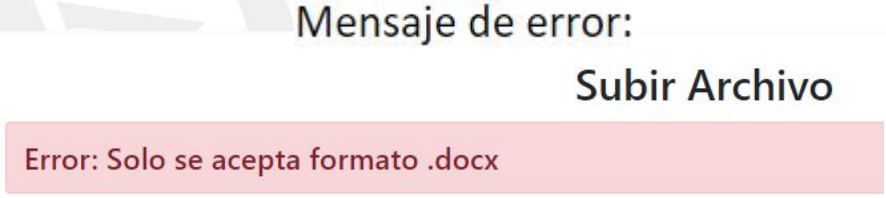
Universidad de Jaén

DESARROLLO

- Resultados Obtenidos (Web):







Universidad de Jaén

PROCESAMIENTOS

Procesamiento 1	ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601EX2AA_2018024087.pdf AT< Instructor: INS< Secretario: 105284
	ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601FX2AA_2016018028.pdf AT< Instructor: INS< Secretario: AT< Dep
	ANONIMIZADO_0500_Caso_C0740401_obj_79053671.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: XXXXX Atestado n°: AT<<<<
	ANONIMIZADO_0500_Caso_C0740406_obj_78906448.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: Atestado n°: AT<<<<< Depe

Universidad de Jaén

PROCESAMIENTOS

Procesamiento 1	ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601EX2AA_2018024087.pdf AT< Instructor: INS< Secretario: 105284 A ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601FX2AA_2016018028.pdf AT< Instructor: INS< Secretario: AT< Depe ANONIMIZADO_0500_Caso_C0740401_obj_79053671.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: XXXXX Atestado n°: AT<<<<< ANONIMIZADO_0500_Caso_C0740406_obj_78906448.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: Atestado n°: AT<<<<< Depen
Procesamiento 2	Instructor Secretario 105284 Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien median Atestado Instructor Secretario Atestado n° Dependencia En calidad de denunciante quien mediante n° y número de soporte acredita s Atestado Instructor Secretario Atestado n° Dependencia En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y

Universidad de Jaén

PROCESAMIENTOS

Procesamiento 1	ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601EX2AA_2018024087.pdf AT< Instructor: INS< Secretario: 105284 A ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601FX2AA_2016018028.pdf AT< Instructor: INS< Secretario: AT< Depe ANONIMIZADO_0500_Caso_C0740401_obj_79053671.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: XXXXX Atestado n°: AT<<<<< ANONIMIZADO_0500_Caso_C0740406_obj_78906448.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: Atestado n°: AT<<<<< Depen
Procesamiento 2	Instructor Secretario 105284 Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien median Atestado Instructor Secretario Atestado n° Dependencia En calidad de denunciante quien mediante n° y número de soporte acredita s Atestado Instructor Secretario Atestado n° Dependencia En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y
Procesamiento 3	el en Vía publica urbana esquina Que ha sido informado de la obligación legal que tiene de decir la verdad de LECr y de la posible re Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien mediante n° En calidad de denunciante quien mediante n° y número de soporte acredita ser ANONpais de nacionalidad mujer nacida en Madrid el dia 2 En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y Secretario arriba mencionados Los funcionarios del Cuerpo

Universidad de Jaén

PROCESAMIENTOS

Procesamiento 1

```
ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601EX2AA_2018024087.pdf AT< Instructor: INS< Secretario: 105284 A
ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601FX2AA_2016018028.pdf AT< Instructor: INS< Secretario: AT< Depe
ANONIMIZADO_0500_Caso_C0740401_obj_79053671.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: XXXXX Atestado n°: AT<<<<
ANONIMIZADO_0500_Caso_C0740406_obj_78906448.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: Atestado n°: AT<<<<< Deper
```

Procesamiento 2

```
Instructor Secretario 105284 Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien
Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien median
Atestado Instructor Secretario Atestado n° Dependencia En calidad de denunciante quien mediante n° y número de soporte acredita s
Atestado Instructor Secretario Atestado n° Dependencia En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y
```

Procesamiento 3

```
el en Vía publica urbana esquina Que ha sido informado de la obligación legal que tiene de decir la verdad de LECr y de la posible re
Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien mediante n°
En calidad de denunciante quien mediante n° y número de soporte acredita ser ANONpais de nacionalidad mujer nacida en Madrid el dia 2
En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y Secretario arriba mencionados Los funcionarios del Cuerpo
```

Procesamiento 4

```
105284 quien extranjero habitación esquina tiene o o o o trabaja ejerciendo prostitución voluntaria esquina desde indicado habi
quien tlfnomóvil robo o intimidación ocurrido señalada paseando estaba escribiendo un mensaje texto particular repentina acerca
quien anonpais 2000 formular psiquicos sufridos fueron causados produjeron 2313 piso une 1999 es exnovio coordinación fuerzas e
huelva 03 00 destinados nacidoa 1978 hijoa domiciliado huelva manifiestanque comparecen 0014 otro o alameda sundheim huelva 001
```

Universidad de Jaén

PROCESAMIENTOS

Procesamiento 1	ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601EX2AA_2018024087.pdf AT< Instructor: INS< Secretario: 105284 / ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P07601FX2AA_2016018028.pdf AT< Instructor: INS< Secretario: AT< Depe ANONIMIZADO_0500_Caso_C0740401_obj_79053671.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: XXXXX Atestado n°: AT<<<< ANONIMIZADO_0500_Caso_C0740406_obj_78906448.PDF Atestado: AT<<<<< Instructor: NUM<< Secretario: Atestado n°: AT<<<<< Deper
Procesamiento 2	Instructor Secretario 105284 Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien median Atestado Instructor Secretario Atestado n° Dependencia En calidad de denunciante quien mediante n° y número de soporte acredita s Atestado Instructor Secretario Atestado n° Dependencia En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y
Procesamiento 3	el en Vía publica urbana esquina Que ha sido informado de la obligación legal que tiene de decir la verdad de LECr y de la posible re Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de quien mediante n° En calidad de denunciante quien mediante n° y número de soporte acredita ser ANONpais de nacionalidad mujer nacida en Madrid el dia 2 En Huelva siendo las 03 horas 00 minutos del día 2019 ante el Instructor y Secretario arriba mencionados Los funcionarios del Cuerpo
Procesamiento 4	105284 quien extranjero habitación esquina tiene o o o o trabaja ejerciendo prostitución voluntaria esquina desde indicado habi quien tlfnomóvil robo o intimidación ocurrido señalada paseando estaba escribiendo un mensaje texto particular repentina acerca quien anonpais 2000 formular psiquicos sufridos fueron causados produjeron 2313 piso une 1999 es exnovio coordinación fuerzas e huelva 03 00 destinados nacidoa 1978 hijoa domiciliado huelva manifiestanque comparecen 0014 otro o alameda sundheim huelva 001
Procesamiento 4 Stop V	105284 extranjero habitación esquina trabaja ejerciendo prostitución voluntaria esquina indicado habiendo problema ninguna trabaja zona eje tlfnomóvil robo intimidación ocurrido señalada paseando escribiendo mensaje texto particular repentina acerca espalda agarra estirando fuer anonpais 2000 formular psiquicos sufridos fueron causados produjeron 2313 piso une 1999 exnovio coordinación fuerzas instrucciones vigentes huelva 03 00 destinados nacidoa 1978 hijoa domiciliado huelva manifiestanque comparecen 0014 alameda sundheim huelva 0014 realizaba tareas

Universidad de Jaén

ANÁLISIS DEL CLASIFICADOR

- Métricas Obtenidas:

	Macro avg F1				
Modelo	Procesamiento 1	Procesamiento 2	Procesamiento 3	Procesamiento 4	amiento 4 Stop
SVM	1	0,99	0,98	0,96	0,93
Naive Bayes	1	0,94	0,97	0,88	0,88
Distilbert	1	0,85	0,91	0,94	0,90
Beto	1	0,99	0,97	0.84	0,83
Maria	1	0,92	0,94	0,85	0,86
Bertin	1	0,73	0,95	0,90	0,84
Multilingual Bert	1	0,43	0.82	0,83	0,86

ANÁLISIS DEL CLASIFICADOR

- Métricas Obtenidas:

Macro avg F1					
Modelo	Procesamiento 1	Procesamiento 2	Procesamiento 3	Procesamiento 4	amiento 4 Stop
SVM	1	0,99	0,98	0,96	0,93
Naive Bayes	1	0,94	0,97	0,88	0,88
Distilbert	1	0,85	0,91	0,94	0,90
Beto	1	0,99	0,97	0.84	0,83
Maria	1	0,92	0,94	0,85	0,86
Bertin	1	0,73	0,95	0,90	0,84
Multilingual Bert	1	0,43	0.82	0,83	0,86

- El Procesamiento 1 “hace trampa”. Se aprovecha de la primera línea -> **Predicción irreal**.

Primera línea en denuncias de Odio

ANONIMIZADO_ANONIMIZADO_dataset/pdfs_denuncias/150_AT/P02023AX2AA_2016006223.pdf

Primera línea en denuncias de No Odio

ANONIMIZADO_0500_Caso_C0740394_obj_78677408.PDF

ANÁLISIS DEL CLASIFICADOR

- Métricas Obtenidas:

Macro avg F1					
Modelo	Procesamiento 1	Procesamiento 2	Procesamiento 3	Procesamiento 4	amiento 4 Stop
SVM	1	0,99	0,98	0,96	0,93
Naive Bayes	1	0,94	0,97	0,88	0,88
Distilbert	1	0,85	0,91	0,94	0,90
Beto	1	0,99	0,97	0.84	0,83
Maria	1	0,92	0,94	0,85	0,86
Bertin	1	0,73	0,95	0,90	0,84
Multilingual Bert	1	0,43	0.82	0,83	0,86

- La estructura de cada tipo de denuncia está **muy diferenciada** -> Resultados de alto rendimiento:

```
Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de qui
Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de qui
Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de qui
Instructor Secretario Dependencia En siendo las del ante el Instructor y Secretario arriba mencionados En calidad de qui
Atestado Instructor Secretario Atestado n° Dependencia Se extiende a las 02 horas 03 minutos del día 2019 para hacer con
Atestado Instructor Secretario Atestado n° Dependencia En Huelva siendo las 01 horas 07 minutos del día 2019 ante el Ins
Atestado Instructor Secretario Atestado n° Dependencia Se extiende en siendo las 02 horas 39 minutos del dia 2019 por el
Atestado Instructor Secretario Atestado n° Dependencia En Sevilla siendo las 02 horas 57 minutos del dia 2019 ante el In
```

Universidad de Jaén

MODELOS NO BERT ENTRENADOS

- *Naive-75deCada:*
  - Modelo **Naive Bayes.**
  - Entrenamiento: **75 primeras** denuncias de Odio y **75 primeras** de No Odio.

- *SVM-75deCada:*
  - Modelo **SVM.**
  - Entrenamiento: **75 primeras** denuncias de Odio y **75 primeras** de No Odio.

Universidad de Jaén

MODELOS BERT ENTRENADOS

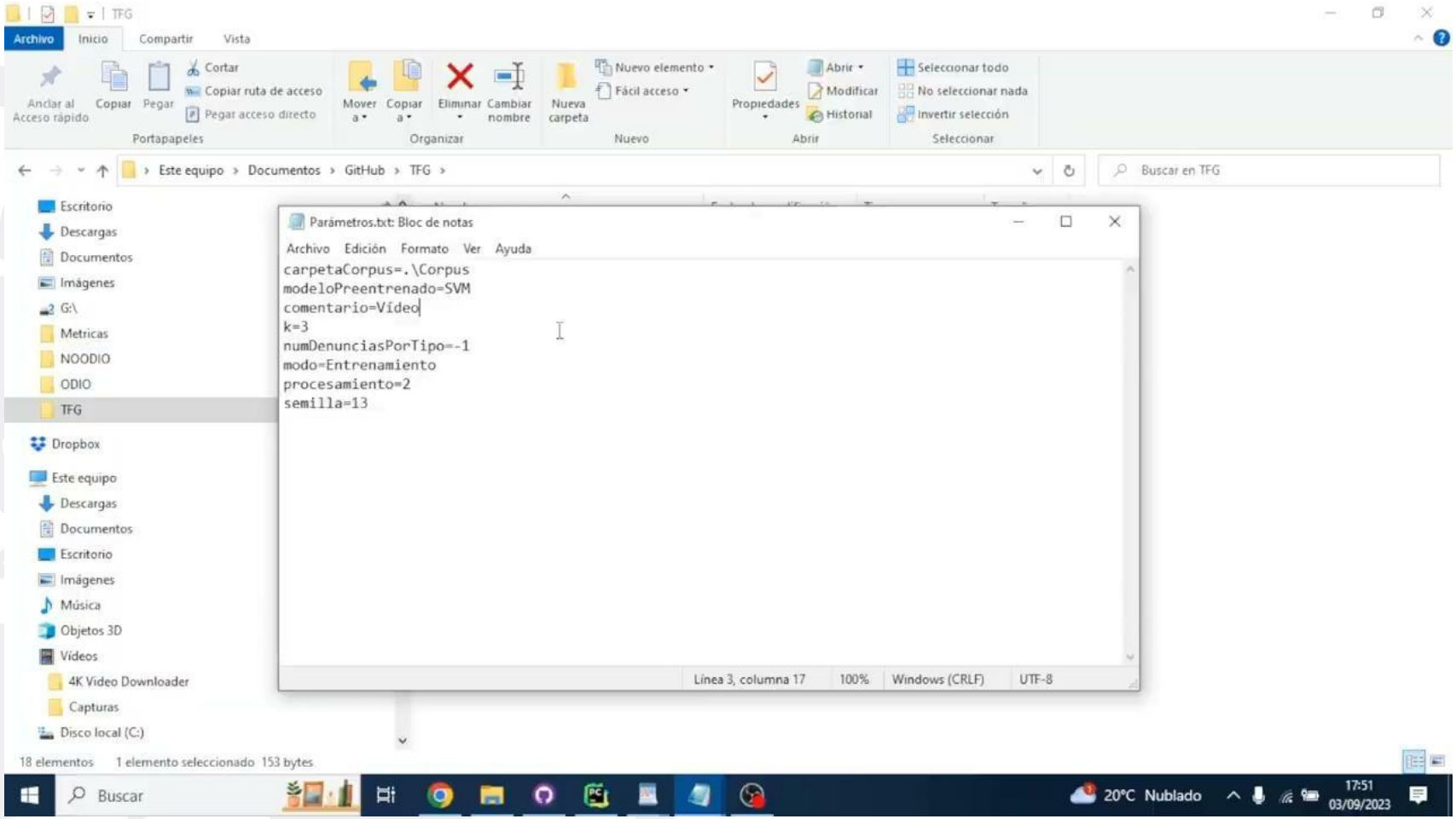
- *dccuchilebert-base-spanish-wwm-cased_75DeCada:*
  - Modelo **BETO.**
  - Entrenamiento: **75 primeras** denuncias de Odio y **75 primeras** de No Odio.

- *distilbert-base-uncased_75deCada:*
  - Modelo **BERT reducido** con palabras en minúscula.
  - Entrenamiento: **75 primeras** denuncias de Odio y **75 primeras** de No Odio.

- *PlanTL-GOB-ESroberta-base-bne_CorpusSubcojuntoKFold:*
  - Modelo **BERTIN.**
  - Entrenado como en la **última división k-fold** de la validación cruzada con la **semilla 42**.
    - Entrenamiento: 127 denuncias (55 No Odio y 72 de Odio).
    - Validación: 63 denuncias (38 de Odio y 25 de No Odio).

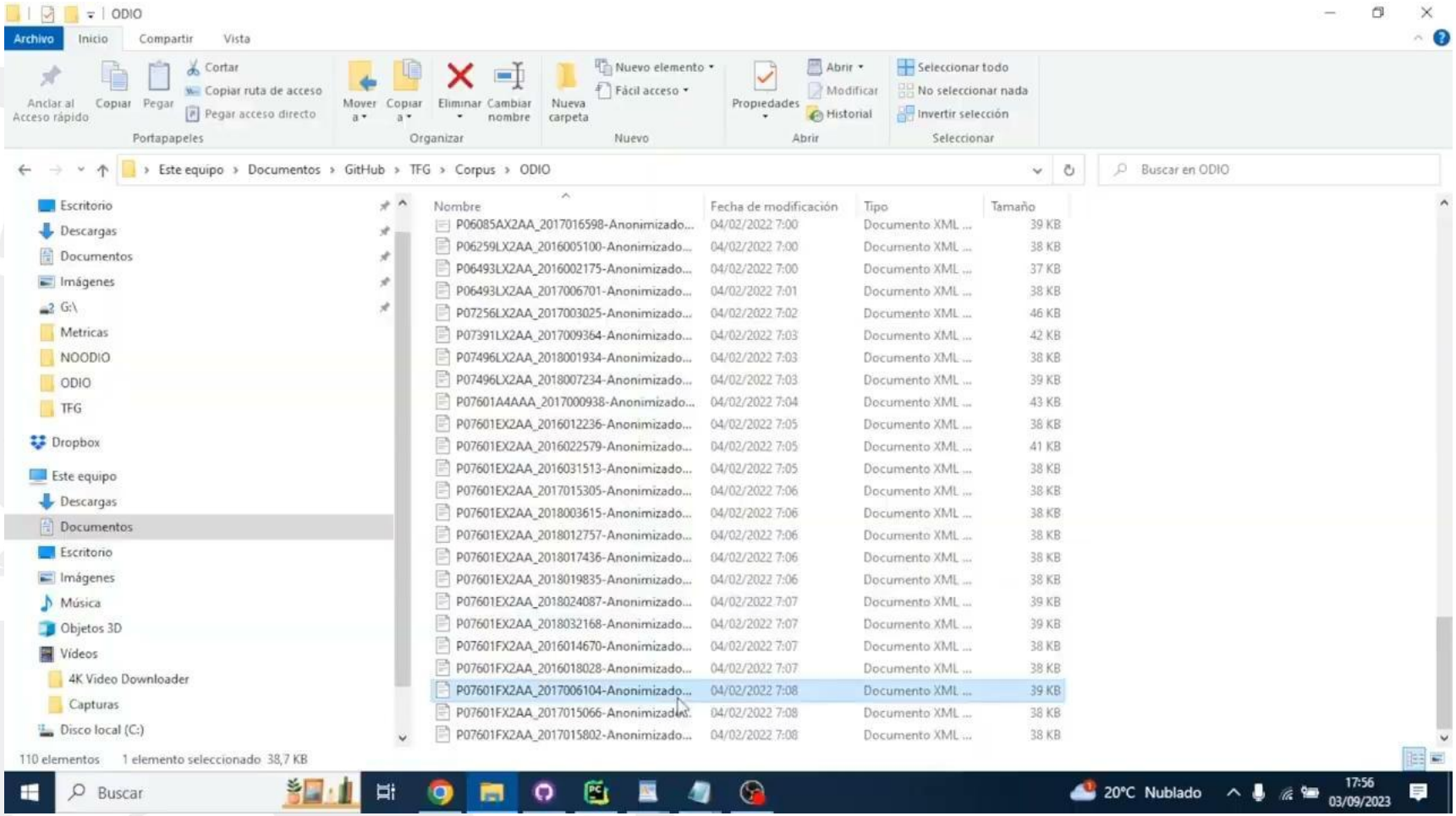
Universidad de Jaén

DEMO: EJECUCIÓN DEL CLASIFICADOR



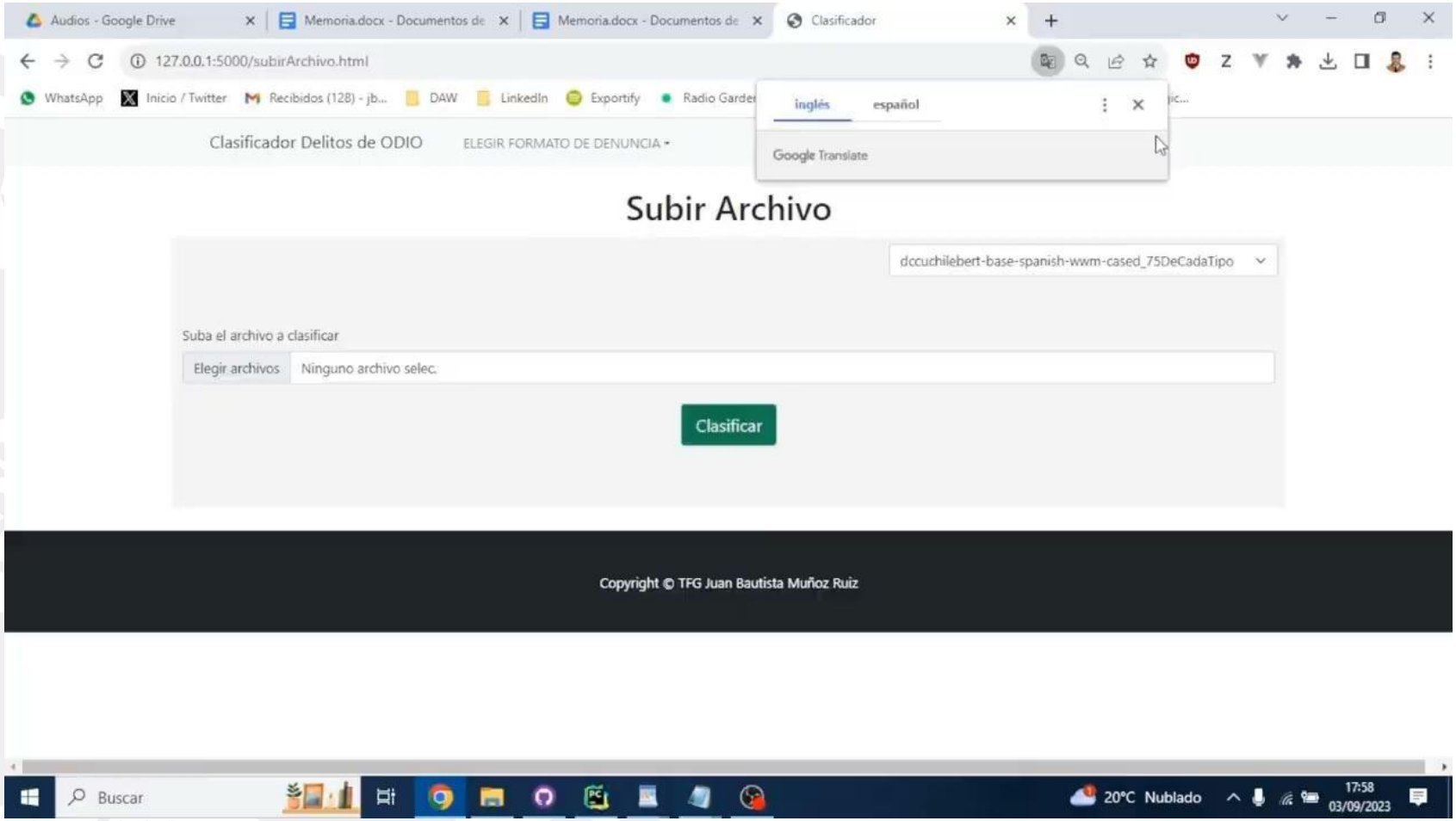
Universidad de Jaén

DEMO: DENUNCIA EN FORMATO ESCRITO



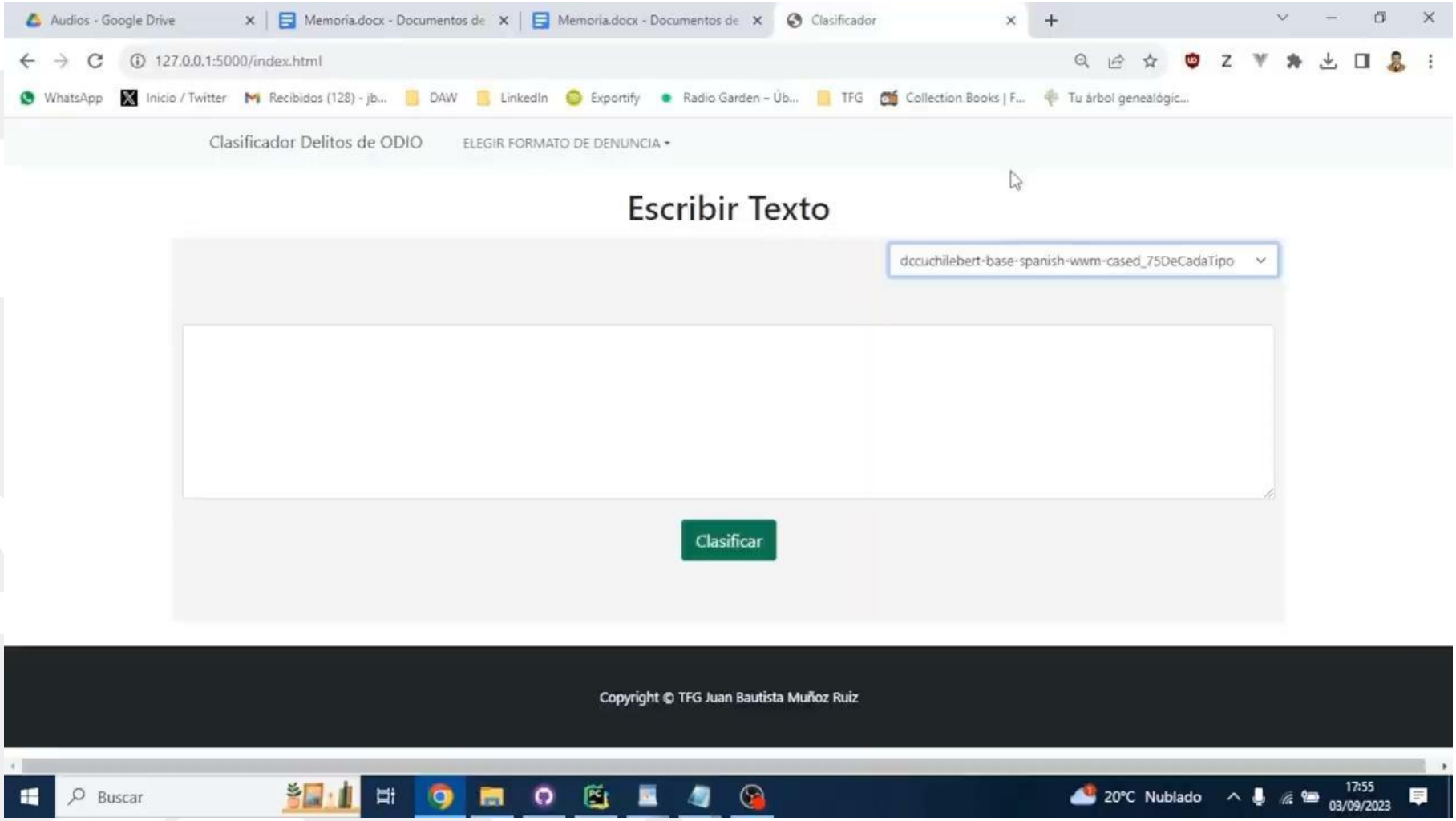
Universidad de Jaén

DEMO: DENUNCIA EN FORMATO ARCHIVO



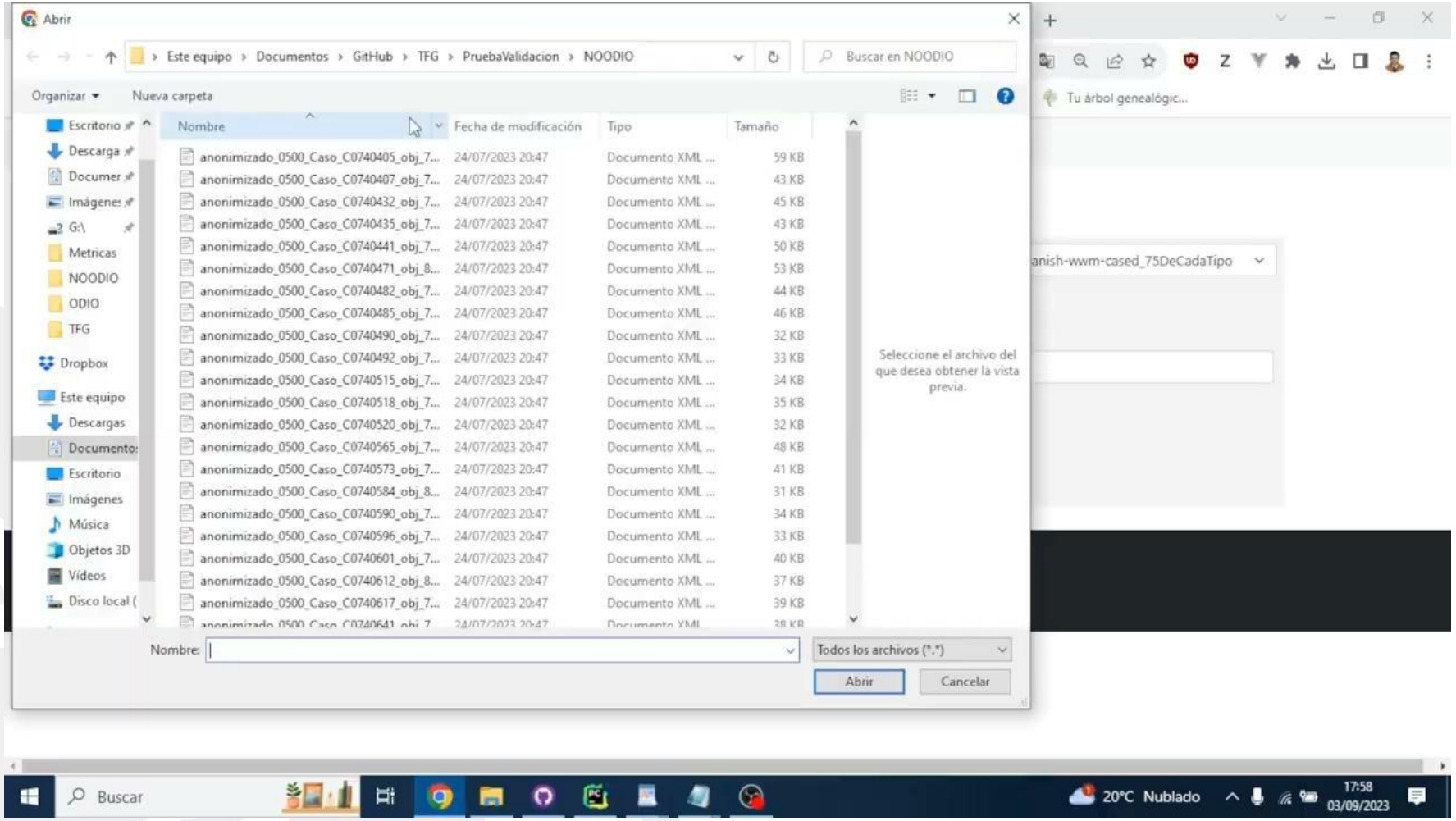
Universidad de Jaén

DEMO: SELECCIÓN DE MODELOS



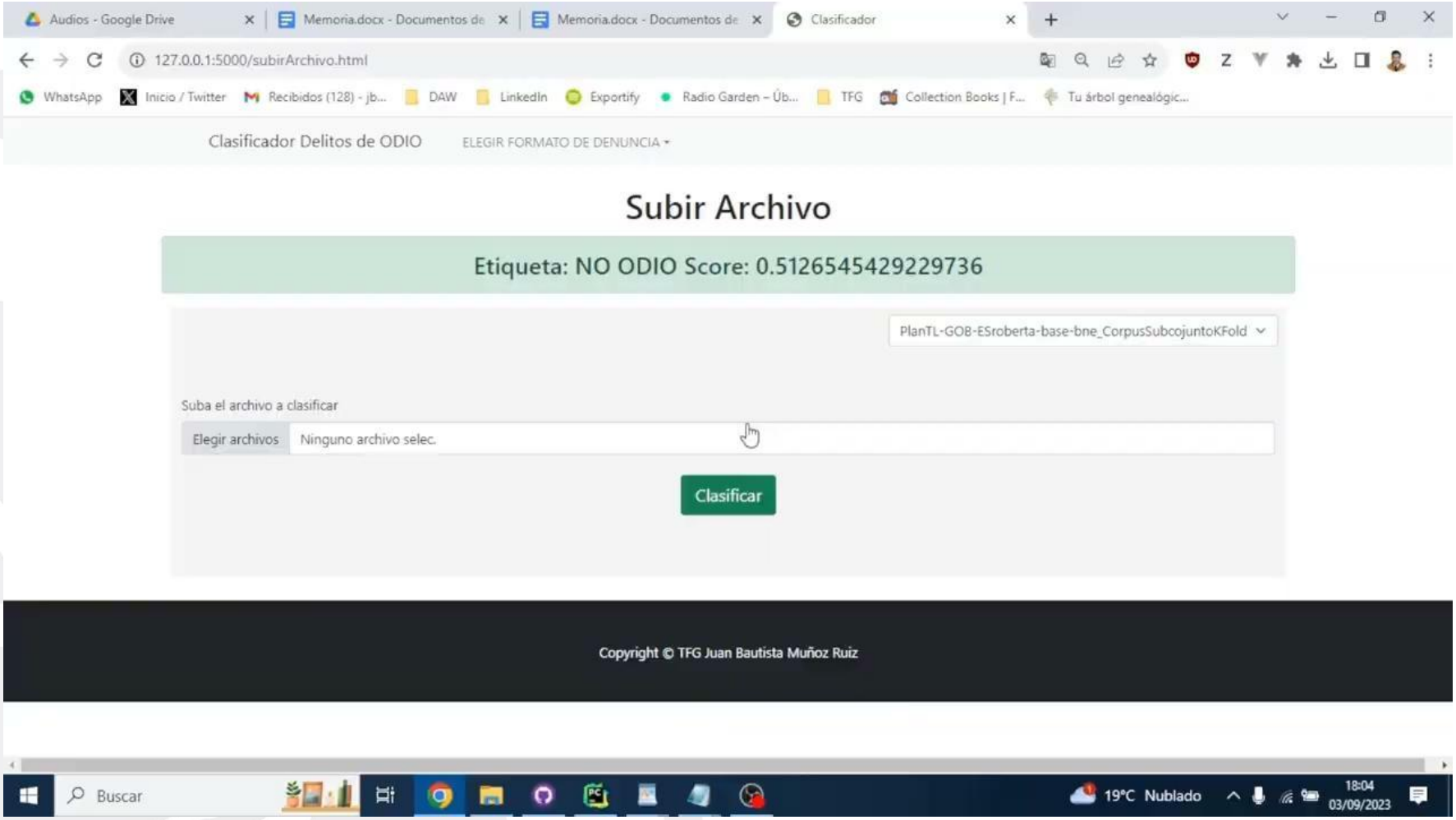
Universidad de Jaén

DEMO: VARIAS CLASIFICACIONES



Universidad de Jaén

DEMO: PREDICCIÓN FALLIDA



Universidad de Jaén

CONCLUSIONES

- El sistema **cumple su objetivo**.
- **Web intuitiva** y fácil de utilizar.
- Se han **aprendido** y afianzado gran variedad de **conceptos**.

- Aspectos a mejorar y **trabajos futuros**:
  - Mayor complejidad en los servicios web.
  - Alojamiento web en un servidor.
  - Procesamientos más sofisticados.
  - Interfaz para el clasificador.

Universidad de Jaén

¡MUCHAS GRACIAS POR SU ATENCIÓN!

"El único límite de la Inteligencia Artificial es la imaginación humana"- **Chris Duffey**