



UNIVERSIDAD
COMPLUTENSE
MADRID



Ejercicios de evaluación

MODULE / COURSE

Máster data science, big data & business analytics

TOPIC/UNIT

Text Mining

PROFESSOR:

Luis Gascó Sánchez, PhD.



Ejercicio práctico (Tiempo estimado de realización: ~ 4-6 horas)

El propósito de este ejercicio es aplicar los conocimientos adquiridos sobre análisis exploratorio de datos textuales y entrenamiento de modelos supervisados, utilizando datos reales extraídos de redes sociales.

Trabajarás con datos de la **shared-task ProfNER**, específicamente con los de la Tarea 1, que se centra en la **detección de profesiones** mencionadas en textos publicados en **Twitter durante la pandemia de COVID-19**. Con esos datos, deberás desarrollar un **clasificador binario** que, dado un tweet, determine si **menciona al menos una profesión (etiqueta 1) o no (etiqueta 0)**. Esta tarea permite identificar contenido relacionado con ocupaciones, con el objetivo de analizar qué profesiones podrían haber sido más vulnerables durante la crisis sanitaria.

Recursos proporcionados

- Un [notebook de Google Colab](#) con la estructura básica del ejercicio. Incluye funciones y celdas predefinidas **que no deben ser modificadas**.
- Un **conjunto de datos** en formato dataset de Hugging Face, que contiene:
 - **Datos de entrenamiento:** con identificadores, texto y etiquetas.
 - **Datos de validación:** con la misma estructura que el conjunto de entrenamiento.
 - **Datos de prueba (test):** solo incluye identificadores y texto (sin etiquetas)¹.

Tareas a llevar a cabo:

1. Análisis Exploratorio de Datos:

- Calcular estadísticas básicas: número de documentos, distribución de clases, longitud de los textos, etc.
- Crear visualizaciones (por ejemplo, wordclouds) para entender mejor el contenido del dataset.

2. Selección y justificación del modelo

- Elegir un modelo preentrenado de HuggingFace adecuado para el idioma y la naturaleza de los datos.
- Justificar claramente por qué se ha seleccionado ese modelo.

3. Entrenamiento del clasificador.

- Entrenar el modelo de forma reproducible.
- Evaluar su rendimiento sobre el conjunto de validación.

4. Generación de predicciones sobre el test

- Aplicar el modelo entrenado sobre el conjunto de test.
- Guardar las predicciones en un archivo -tsv con el siguiente formato:
 - Dos columnas **id** y **label**, separadas por tabulador.
 - El archivo debe llamarse

APELLIDO1_APELLIDO2_NOMBRE_ejercicio1_predicciones.tsv.

¹ El test tiene campo de etiquetas, pero todas tienen el valor -1.

Instrucciones de entrega

Debes entregar los siguientes archivos:

1. **Notebook de Google Colab** con el desarrollo completo del ejercicio, incluyendo análisis, código y justificaciones. El nombre del archivo debe seguir esta nomenclatura:

APELLIDO1_APELLIDO2_NOMBRE_ejercicio1.ipynb

2. **Archivo de predicciones** generado por el modelo, con el nombre:

APELLIDO1_APELLIDO2_NOMBRE_ejercicio1_predicciones.tsv

Ambos archivos deben comprimirse en un archivo **.zip** con nombre:

APELLIDO1_APELLIDO2_NOMBRE_TM_EJERCICIO.zip

Criterios de evaluación

Criterio	Peso
Análisis exploratorio y preprocesamiento	20%
Selección y justificación del modelo	25%
Formato y validez del archivo de predicciones	5%
Ejecución correcta del notebook sin intervención (sin errores)	10%
Rendimiento del modelo en el conjunto de prueba	30%
Claridad y calidad de las explicaciones	10%

Nota importante: El rendimiento del modelo se evaluará utilizando métricas estándar como el **F1-score** sobre el conjunto de test. Si el archivo de predicciones **no cumple con el formato requerido**, el ejercicio **no podrá ser evaluado**. El primer clasificado de la clase obtendrá la máxima calificación en este apartado, disminuyendo de forma proporcional al nivel de f1-score obtenido.

Ejecutabilidad: El notebook debe poder ejecutarse **de principio a fin sin intervención manual**. Si utilizas librerías no incluidas en el entorno por defecto de Google Colab, **debes instalarlas en el propio notebook**. Para que pueda ejecutarse de principio a fin

Recomendaciones

- Trabajar directamente en el entorno de Google Colab para evitar problemas de compatibilidad o instalación de librerías.
- Sigue la estructura del notebook y **no modifiques las celdas indicadas como fijas**.
- Documentar adecuadamente cada paso realizado, incluyendo justificaciones para las decisiones tomadas.
- Realizar pruebas con subconjuntos de datos antes de entrenar el modelo completo para verificar que todo funciona correctamente.