Conjunto de datos

Para esta actividad se utilizará un conjunto de datos etiquetados para la detección de discursos de odio y lenguaje ofensivo en tweets en inglés. El dataset está contenido en el archivo **tweets_hate_speech.csv** y almacena las siguientes variables:

- tweet – Texto del tweet.
- label – Etiqueta asignada (*hate, offensive, neither*).

Práctica

Utilizando Python, sin recurrir a otros paquetes externos que no sean los vistos en el curso:

1) Realizar un análisis exploratorio de datos utilizando distintos tipos de funciones, gráficas y medidas estadísticas. Presentar los resultados más importantes.
2) Preprocesar el conjunto de datos para que pueda utilizarse en modelos de aprendizaje automático teniendo en cuenta:
    - Preprocesamiento de texto.
    - Procesamiento de variables de salida.
    - Representaciones vectoriales del lenguaje.
    - Balance del conjunto de datos.
3) Entrenar y mostrar los resultados obtenidos con al menos tres modelos vistos en el curso, teniendo en cuenta:
    - División del conjunto de datos para entrenamiento y prueba.
    - Búsqueda de hiperparámetros para mejorar el rendimiento de los modelos.
    - Evaluación del modelo con las métricas convenientes.

Entrega

1) Un archivo .ipynb conteniendo la solución a los incisos 1, 2 y 3.
2) Un informe que documente y justifique los distintos pasos realizados, herramientas y técnicas utilizadas.
3) El trabajo se defenderá en un coloquio en fecha y hora a publicar.