Caso grupal: clasificación de mensajes de odio

Objetivos

Mediante este trabajo se pretende que el estudiante ponga en práctica la aplicación de un proceso completo de *machine learning* a la clasificación de mensajes de odio. Por tanto, debe detallar los pasos que hay que realizar para la clasificación de mensajes, así como una evaluación de los resultados:

* Decidir la técnica de aprendizaje automático a utilizar.
* Analizar diferentes modelos de aprendizaje supervisado para la clasificación de los mensajes de odio.
* Evaluar los resultados obtenidos por cada modelo en términos de precisión, *recall,* y *f1.*

Pautas de elaboración

La empresa SureTech Innovations quiere implementar un clasificador de mensajes de odio. Para ello, se deben entrenar diferentes modelos de clasificación, evaluar el rendimiento de cada uno y, finalmente, decidir cuál sería el mejor.

Para llevar a cabo esta actividad cuentas con el siguiente *dataset*:

Said-Hung, E., Römer Pieretti, M., Montero-Diaz, J., De Lucas Vicente, A. y Torres, J. M. (2023). *Hate Speech Library in Spanish. Github.* <https://github.com/esaidh266/Hate-Speech-Library-in-Spanish/tree/main>

Los pasos que se deben seguir para desarrollar esta actividad son:

1. ¿Qué tarea de aprendizaje automático se pretende resolver? ¿Qué análisis previos serían necesarios realizar para saber si es posible construir el modelo que buscamos? Realiza un análisis exploratorio de datos y busca que sea adecuado para la tarea que se pretende resolver y para el *dataset* facilitado. Explica con detalle los análisis que serían necesarios realizar sobre los datos.
2. En base a la tarea que se pide realizar y a las características del *dataset* decide qué algoritmo o algoritmos se deben utilizar. Implementa los modelos necesarios. Realiza los entrenamientos necesarios para cada uno de ellos (al menos se deben hacer entrenamientos con dos algoritmos diferentes). Explica por qué son convenientes los algoritmos elegidos.
3. Realiza una explicación de las métricas de evaluación que se deben utilizar para valorar la bondad del modelo. Justifica la decisión de las elegidas. ¿Cómo habría que evaluar el modelo? Evalúa los modelos entrenados en el punto anterior y explica los resultados obtenidos.

Extensión y formato

Notebook Python.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Caso grupal: clasificación de mensajes de odio | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Realiza correctamente el análisis de datos | 1,5 | 15% |
| Criterio 2 | Realiza correctamente la implementación de los modelos de entrenamiento | 3,5 | 35% |
| Criterio 3 | Realiza correctamente la evaluación de los modelos entrenados | 2,5 | 25% |
| Criterio 4 | Justifica y responde de manera razonada todas las preguntas planteadas | 2,5 | 25% |
|  |  | **10** | **100 %** |