Master of Science in Data Science

Métodos de aprendizaje de máquinas en Data Science

Tarea 3: Modelos de Clasificación

El reconocimiento de discursos de odio, o Hate Speech Recognition en inglés, es una tarea crucial en el campo de la inteligencia artificial debido a su importancia en la promoción de un entorno en línea seguro, un ambiente más saludable y en la protección de derechos fundamentales de los seres humanos. Se considera como discurso de odio los ataques verbales o escritos por motivos de raza, religión, género u orientación sexual, que crean un ambiente tóxico que puede tener consecuencias psicológicas y emocionales devastadoras. Identificar y eliminar este tipo de contenido, es un precedente fundamental para un entorno más seguro y amigable.

Es bien sabido que las redes sociales son lugares propicios para emitir estos discursos debido al anonimato, el gran alcance, facilidad de acceso y la imposibilidad de recibir represalias. Plataformas como Twitter/X han mostrado gran preocupación por erradicar este tipo de comportamiento. Lamentablemente la mayoría de los esfuerzos para poder combatir este tipo de problemática se hace en inglés, lo cual hace que muchas de las herramientas existentes no den los resultados esperados en nuestro idioma. Además, como caso particular del Español, el "Chileno" es una variante aún más difícil dado la cantidad de modismos con los que hablamos habitualmente.

Dada la problemática en cuestión, se cuenta con un set de Tweets en español "chileno" con el cuál se busca crear una herramienta para detectar Odio dentro del contenido de un Tweet. Para ello se le hará entrega de un conjunto de Tweets con el que deberá entrenar un modelo de Machine Learning capaz de clasificar el Tweet como Odio o no. Además, se le hará entrega de un Set de Test el cuál no posee etiquetas y deberá clasificarlos para poner a prueba su modelo entrenado, en formato competencia.

El set de datos de entrenamiento contiene 2,256 Tweets incluyendo un tweet_id, el contenido del Tweet, y una etiqueta indicando el número de Anotadores que considera que el Tweet contiene Odio.

Para la evaluación de su algoritmo, se entrega un set de datos de test con 2,291 Tweets que sólo contienen el tweet_id y el contenido del Tweet.

Cabe destacar que al igual que en la tarea anterior, los modelos se pondrán a competir y se evaluarán en función de su rendimiento de acuerdo al **F1-Score**, usando como clase positiva la clase odio.

Para esta entrega usted deberá entregar 2 archivos

1. Un archivo ipynb que muestre todo el proceso de preprocesamiento y limpieza de datos aplicados. Además, deberá mostrar el modelo utilizado y la búsqueda de hiperparámetros respectiva. Este archivo ya deberá haber sido ejecutado y cuando se cargué uno deberá ver todo el proceso de ejecución.

2. Un archivo csv conteniendo los 2,291 tweet_id del set de Test y la predicción asociada (1 o 0).

La entrega a través de GitHub tendrá una bonificación de 5 décimas.

Los puntajes asignados a cada tarea corresponden a:

- Limpieza de datos y preprocesamiento (1.5 puntos).
- Aplicación de modelo de SVM (1 punto).
- Aplicación de modelo de Naive Bayes (1 punto).
- Aplicación de modelo de ensamblado (1 punto).
- Comparación y selección del mejor modelo (0.5 puntos).
- Competencia (1 punto).

La fecha de entrega es el 10 de diciembre a las 23:59 horas.
La tarea se puede realizar hasta en grupos de 2 personas.
¡Mucha suerte!

Diccionario de variables:

- **tweet_id**: Un identificador numérico del Tweet.
- **text**: Contenido del Tweet como Texto libre.
- **Odio**: Número de anotadores que consideran que el Tweet contiene Odio o no.