uc3m | Universidad Carlos III de Madrid

OpenCourseWare

Procesamiento de Lenguaje Natural con Aprendizaje Profundo,

Práctica: SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS)

El objetivo de esta práctica es que apliques todo lo aprendido durante el curso para desarrollar diferentes enfoques basados en aprendizaje profundo para abordar las tareas propuestas en la competición SemEval-2023 Task 10, cuyo principal objetivo era la detección de seximos en mensajes de redes sociales. Es una tarea similar a la que ya hemos visto en los ejercicios (EXIST), aunque tiene un dataset distinto y la clasificación de los mensajes sexistas también es distinta.

En esta práctica deberás tratar de abordar todas las fases necesarias para el desarrollo y evaluación de modelos basados en aprendizaje profundo para resolver las tres tareas propuestas en la competición.

El sexismo es un problema creciente, especialmente, en las redes sociales. Hoy en día existen herramientas automatizadas que permiten identificar contenidos sexistas, pero la mayoría sólo proporciona clasificaciones para categorías genéricas de alto nivel, sin proporcionar ningún tipo de explicación. La competición SemEval-2023 taks 10 persigue fomentar la investigación en el desarrollo de modelos en inglés para la detección del sexismo que sean más precisos y explicables, con clasificaciones detalladas del contenido sexista en redes sociales como Gab y Reddit.

La competición propone tres subtareas:

- TAREA A - Detección de sexismo binario: clasificación binaria donde los sistemas tienen que predecir si una publicación es sexista o no.
- TAREA B - Categoría de sexismo: para mensajes sexistas, una clasificación de cuatro clases donde los sistemas tienen que predecir una de las siguiente cuatro categorías: (1) amenazas, (2) derogación, (3) animosidad, (4) discusión prejuiciosa.
- TAREA C - Vector de sexismo detallado: también únicamente para mensajes sexistas, consiste en una clasificación de 11 clases donde los sistemas tienen que predecir uno de los 11 clases. Puedes ver esta clasificación en el siguiente enlace: https://github.com/rewire-online/edos/blob/main/edos_vectors.png

Los organizadores de la competición proporcionaron un dataset, que se puede

descargar desde https://github.com/rewire-online/edos/tree/main/data. En concreto, el dataset consta de 20.000 instancias o mensajes (10.000 son de Gab y 10.000 de Reddit).

A continuación, presentamos brevemente las redes sociales Gab y Reddit. Gab es un servicio estadounidense de redes sociales y microblogging de tecnología alternativa conocido por su base de usuarios de extrema derecha. Gab ha atraído a usuarios y grupos que han sido excluidos de otras plataformas de redes sociales y a usuarios que buscan alternativas a las principales plataformas de redes sociales. Por su parte, Reddit es un sitio web estadounidense de agregación de noticias sociales, clasificación de contenidos y debates. Los administradores de Reddit moderan las comunidades. La moderación también la llevan a cabo moderadores específicos de la comunidad, que no son empleados de Reddit.

Cada instancia del dataset contiene los siguientes campos o columnas:

- rewire_id: un identificador único para cada entrada
- texto: el texto del mensaje, es decir, el texto de entrada
- label_sexist: etiqueta de tarea A
- label_category: etiqueta de la tarea B.
- label_vector: etiqueta de la tarea C.

Las clases de los campos label_category y label_vector se pueden consultar en la página web de la tarea:
https://github.com/rewire-online/edos/blob/main/edos_vectors.png

NOTA: Para las instancias no sexistas, label_category y label_vector tienen el valor None.

En esta práctica se pide:

1) Estudiar el dataset, en particular, su distribución de clases (para cada tarea) y la distribución de los tamaños de sus textos.
2) Ajusta distintos transformers para cada una de las tareas (A, B y C), y analiza sus resultados sobre el conjunto test. Para cada modelo y tarea, almacena sus resultados en un fichero, para analizarlos y determina cuál es el mejor modelo para cada tarea.
3) Aplicando las técnicas de DA estudiadas durante el curso, genera nuevos ejemplos en el conjunto de entrenamiento y usalos para ajustar los modelos. ¿Las técnicas ayudan a mejorar los resultados?, ¿cuál es la mejor técnica?, ¿Cuántos nuevos textos son necesarios para incrementar los resultados?.