Reporte y Conclusiones Regresión Logística

Mtra. Olivia Gutú
José Benjamín Partida Peralta
José Rodrigo Calles Villaescusa

10/marzo/2022

Se tiene como objetivo de este proyecto crear un modelo de regresión logística para analizar el sentimiento de distintos tweets utilizando un conjunto de datos de tweets relacionados con el presidente de México.

Primero importamos el conjunto de datos que se nos proporcionó, del cuál eliminamos una columna donde los tweets ya estaban limpios, ya que buscábamos hacer nuestra propia limpieza de los datos.

Limpiamos los datos eliminando aspectos como hashtags, menciones, caracteres que no fueran letras en español y también eliminando los acentos para facilitar un poco mas el modelado. Después separamos los datos en entrenamiento y prueba usando la función de train_test_split de sklearn.

Luego usando la función process_tweet() de la utils.py que utilizamos en las libretas anteriores, se tokeniza el tweet en palabras individuales y se eliminan las stop words para poder extraer las características de los tweets. Se extraen las características de una lista de tweets y se guardan en una matriz. La primera característica es el número de palabras positivas en un tweet y la segunda es el número de palabras negativas en un tweet.

Después de hace la función de extracción de características, entrenamos un modelo de regresión logística usando la librería de sklearn e hicimos nuestras predicciones, además de obtener el score que fue de 0.48 en entrenamiento y 0.45 en el conjunto de prueba.

Los scores obtenidos fueron bastantes bajos, probablemente hay que hacer una mejor limpieza de los datos para obtener mejores resultados. También se podría jugar un poco más con el modelo y aplicar distintas técnicas para ajustar los híper parámetros.