Graphical user interface, text, application

Description automatically generated

Escenario 1:  
Dado un corpus con tweets etiquetados como agresivos y no agresivos, crea un clasificador que pueda indicar si un tweet no visto antes contiene mensajes agresivos.

En este caso, utilizaría BoW con un pesado binario, debido a que será capaz de predecir de forma correcta, ya que analiza qué palabras tienden a coexistir en un tweet para convertirlo en algo agresivo.

No utilizaría DOR porque este se desarrolla mejor a la hora de entender contextos por medio de las palabras, sin embargo, dado este problema algunas palabras consideradas como groserías podrían estar en ambos tipos de tweet, agresivos y no agresivos, así que la información que provee no es tan útil.

Escenario 2:

Dado un corpus con textos provenientes de México y Colombia y etiquetados con su respectivo lugar de origen, crea un clasificador que pueda decir de qué país proviene un texto no antes visto.

Para este escenario me iría por DOR, debido a que contiene información útil respecto al contexto en el que se desarrollan los documentos, y sería capaz de encontrar la relación entre una etiqueta y la existencia de específicos términos únicos a ese lugar, además de que su primer término (antes de transformar) permite no dar valores muy grandes en cuestión de frecuencia.

Por otro lado, no escogería BoW porque puede ser posible que malinterprete los pesados con palabras frecuentes en ambos idiomas, y el primer término (en caso de TFIDF) podría disparar los resultados, haciendo que preste más atención a términos como “la” o “el”.

254

DOR y BOW son dos tipos de pesado distintos, entre sus diferencias se encuentra: el tamaño de matrices que generan, BOW provee una matriz de (# docs x # terms), mientras que DOR brinda (# terms x # docs) y eventualmente (# docs x # docs); otra de las diferencias es que BoW propone tener una representación de documentos por medio de una bolsa de palabras, pero DOR busca tomar en cuenta contextos en común de las palabras, utilizando una bolsa de conceptos; y esa última yo creo que es la diferencia más importante, ya que DOR nos permite conocer (tras transformar) la correlación entre documentos, permitiendo tener un valor alto si estos desarrollan contextos similares y es muy útil para su análisis, aunque quizá no tanto para un clasificador clásico, mientras que la BoW tiene una gran pérdida de información, pero aún así brinda buenos resultados a la hora de clasificar.