Análisis de predictores para modelo de aprendizaje supervisado

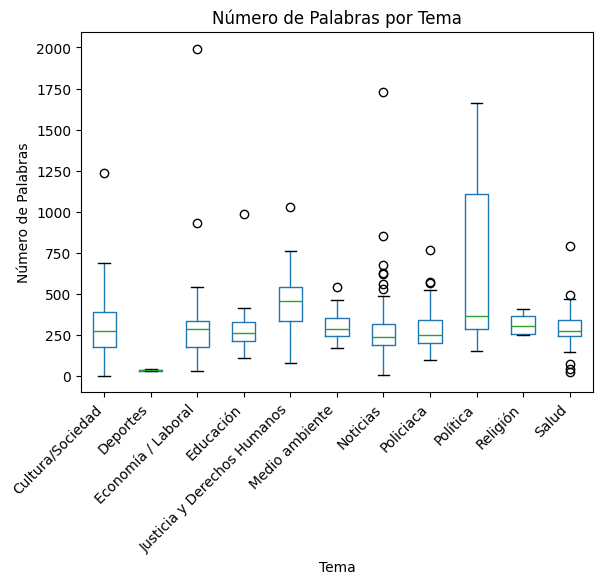
**El objetivo** **de este análisis** es determinar cuáles variables pueden servir de predictores para entrenar un modelo de aprendizaje supervisado que clasifique los textos de varios periodistas en temas dados.

Sobre la estructura de la base de datos.

Tenemos un archivo de Excel con las columnas: 'Periodista', 'Posicion', 'Formato', 'Archivo', 'Tema 1', 'Tema 2', 'Sección', 'Etiquetas', 'Título', 'Medio', 'FDP (aaaa-mm-dd)', 'Enlace', 'Wayback Machine', 'archive.today', 'Captura PNG', 'Página descargada', 'Google Drive', 'Texto', 'Localidad', 'Municipio', 'Entidad', 'CVE\_INEGI', 'CVE\_ENT', 'CVE\_MUN', 'CVE\_LOC', 'Compilación', 'Coautores', 'Notas', 'Duración del video (hh:mm:ss)'. La variable respuesta es ‘Tema1’ mientras que el resto de las variables pueden ser usadas como predictores por si mismas o como herramientas para construir un predictor.

Sobre ‘Texto’

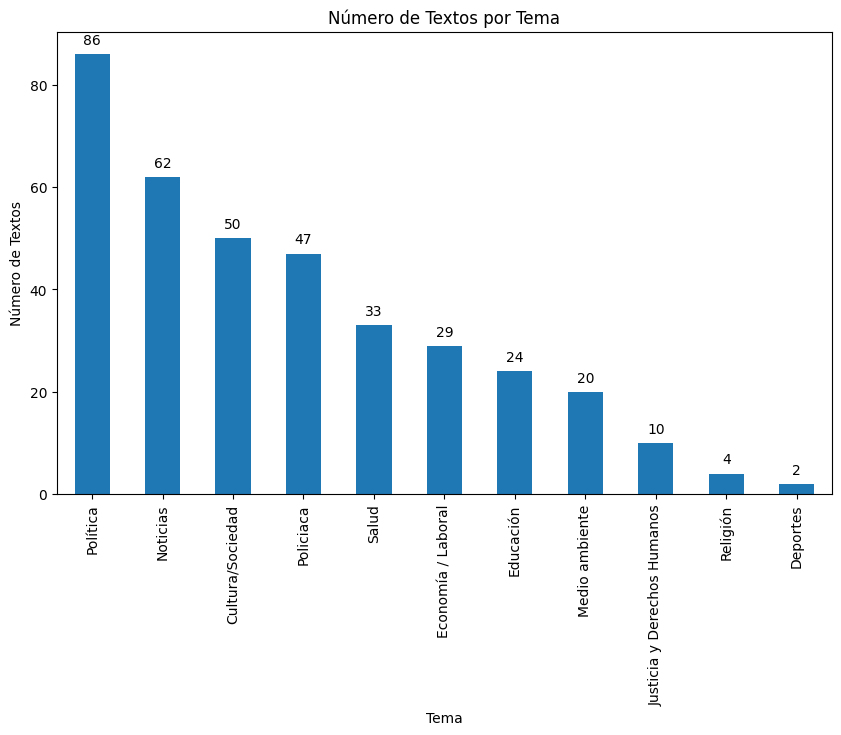
El texto es una de las variables más utiles por que tiene mucha información por si solo. Queremos entender cuales son las diferencias en el texto que determinan la categoría en la que se debe clasificar.



Palabras clave

Menciones de partidos políticos (Morena, PAN, PRI, PT, PRD)

Sobre los limitantes de la muestra



Es importante tomar en cuenta que la base de datos con la que estamos trabajando tiene características que restringen sus usos para el modelo que estamos diseñando. Una de las más importante es la distribución por temas de los textos. A pesar de que temas como política y cultura tienen más de 50 textos cada una, deportes y religión cuentan con menos de 5. Debido a esto y bajo la suposición de que los temas con menor número de textos no son la prioridad de este análisis, sugiero no trabajar con ellos para el entrenamiento o juntarlos creando un nuevo bloque de ‘otros’.

Conclusiones: