Alumna: Melina Markstein

Political social media posts

Este dataset fue obtenido en la página Kaggle:
https://www.kaggle.com/datasets/crowdflower/political-social-media-posts
Este consiste en 5000 publicaciones/posteos de las redes sociales (facebook y twitter) de políticos estadounidense (senadores y representantes del congreso).
Los mensajes fueron clasificados según: audiencia (nacional o local), sesgo (neutral/bipartisan, o biased/partisan) y por el objetivo de los mismos (ataque, apoyo, personal, etc).

El dataset original cuenta con estas columnas:

- _unit_id: un id único para cada mensaje
- golden: siempre FALSE (falso) → La columna fue eliminada
- _unit_state: siempre "finalized" (finalizado) → La columna fue eliminada
- _trusted_judgments: el cantidad de personas que juzgaron el mensaje; un número entero entre 1 y 3
- _last_judgment_at: cuando fue reoclectado el ulitmo "juicio"
- audience: público al que fue dirigido; nacional o circunscripto
- audience:confidence: una medida de confianza en el juicio de la audiencia; un número entre 0,5 y 1 → columna eliminada
- bias: el sesgo que se presenta como neutral o partidario
- bias:confidence: una medida de confianza en el juicio de sesgo; un valor entre 0,5 y 1 → columna eliminada
- message: el objetivo del mensaje:
  - attack: el mensaje ataca a otro político
  - constituency: el mensaje habla de la circunscripción del político
  - information: n mensaje informativo sobre las novedades del gobierno o de los Estados Unidos en general
  - media: un mensaje sobre la interacción con los medios de comunicación
  - Mobilization: mensaje destinado a movilizar a los partidarios.
  - other: una categoría que engloba los mensajes que no encajan en las otras categorías.
  - personal: un mensaje personal, que suele expresar simpatía, apoyo y/o condolencias, u otras opiniones personales
  - policy: un mensaje sobre políticas
  - support: un mensaje de apoyo político
- message:confidence: una medida de confianza en el juicio del mensaje; un valor entre 0,5 y 1 → columna eliminada
- orig_golden: siempre vacío; presumiblemente si alguna parte del mensaje estaba en el estándar de oro → columna eliminada
- audience_gold: siempre vacío; presumiblemente si la respuesta de la audiencia estaba en el patrón oro → columna eliminada
- bias_gold: siempre vacío; presumiblemente, si la respuesta de sesgo estaba en el patrón oro → columna eliminada
- bioid: un id único para el político
- embed: código HTML para incrustar este mensaje → columna eliminada
- id: id único del mensaje DENTRO del sitio de medios sociales del que se extrajo
- label: una frase de la forma "De: nombre apellido (cargo del estado)"

- ➢ message_gold: siempre en blanco; presumiblemente si la respuesta del mensaje estaba en el estándar de oro → columna eliminada
- ➢ source: donde se publicó el mensaje,"facebook" o "twitter"
- ➢ text: el texto del mensaje

Como fue indicado, varias columnas no fueron seleccionadas (eliminadas), ya que estaban relacionadas al proceso de construcción del dataset que no tiene relevancia ahora, siendo que no todas las filas están vacías o todas tienen el mismo valor.

Además se tomó en cuenta un dataset extra:

- ➢ Idiomas: https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

El cual fue recolectado utilizando una extensión de google chrome para pasar la tabla del sitio web a google-sheets.
Este fue utilizado para identificar en qué idioma fueron escritos los posteos con un nombre claro y conocido.

Las nuevas variables que se agregaron fueron:

- ➔ Cantidad_carcteres
- ➔ Language_name
- ➔ Usuario
- ➔ Cargo
- ➔ Estado

Las últimas 3 eran originalmente parte de la columna “label”, sin embargo consideré que sería más enriquecedor analizarlas por separado con respecto a la intención del mensaje.

Otras posibilidades de dataset para conectar:

- ➔ Tomar tabla de regiones de Estados Unidos: https://es.wikipedia.org/wiki/Regiones_de_Estados_Unidos
- ➔ Tomar tabla de longitud y latitud: https://www.latlong.net/category/states-236-14.html

Con estos podría unir los estados a los que pertenecen los políticos a regiones, con el objetivo de ver si hay una tendencia tendiendo en cuenta estos parámetros. Y además sería posible realizar un gráfico teniendo en cuenta las longitudes y latitudes.