Conjunto de problemas 4: Predicción de tweets

“Una rosa con cualquier otro nombre olería igual de dulce”

julieta capuleto

Hay un adagio que dice: “elige bien tus palabras”. Las palabras mismas pueden revelar mucho más de lo que estamos tratando de decir. Cada vez hay más pruebas de que nuestras palabras escritas muestran quiénes somos.

El objetivo es predecir a quién pertenece cada tuit. El conjunto de datos de capacitación contiene tuits de las cuentas de tres destacados políticos colombianos: Claudia López, Gustavo Petro y Álvaro Uribe. El conjunto de datos de prueba contiene 500 tweets sin etiquetar. Queremos predecir qué cuenta publicó los tweets en el conjunto de prueba.

Hay dos resultados esperados:

1. Un documento .pdf.
2. Envíos con los pronósticos de tu equipo en Kaggle en el siguiente enlace.

El documento debe contener las siguientes secciones:

- Introducción. En la introducción se expone brevemente el problema y si existen antecedentes. Describe brevemente los datos y su idoneidad para abordar la pregunta del conjunto de problemas. Contiene una vista previa de los resultados y las conclusiones principales.
- Datos[1] . Al redactar esta sección, debe:
    1. Describa los datos y el proceso de construcción de la muestra, incluido cómo se limpiaron y combinaron los datos y, si se crearon nuevas variables, cómo se crearon.
    2. Presentar un análisis descriptivo/explicativo de los datos usando visualización de texto tecnicas
- Modelo y Resultados. En esta sección se presentan los modelos presentados para evaluación. Al redactar esta sección, incluya:

[1]Esta sección se encuentra aquí para que el lector pueda comprender su trabajo, pero probablemente debería ser la última sección que escriba. ¿Por qué? Porque vas a hacer elecciones de datos en los modelos estimados. Y todas las variables incluidas en estos modelos deben describirse aquí.

  - Una explicación detallada de cómo se entrenó, la selección de hiperparámetros y cualquier otra información relevante.
  - Una comparación con al menos otras 4 especificaciones enviadas a Kaggle.

- Conclusiones y Recomendaciones. En esta sección, expone brevemente las principales conclusiones de su trabajo.

1 Directrices adicionales

- Las predicciones deben enviarse en Kaggle. Consulte el sitio web de la competencia para obtener más información. información.

- Convierte un documento .pdf en Bloque Ne´on. El documento no debe tener más de 8 (ocho) páginas e incluir, como máximo, 8 (ocho) anexos (tablas y/o figuras). La bibliografía y las exhibiciones no cuentan para el límite de páginas. Puede agregar un apéndice, pero el documento principal debe ser independiente. Específicamente, un lector debe poder seguir el análisis en el documento y estar convencido de que es correcto y coherente solo con el texto principal, sin consultar el apéndice.

- El documento debe incluir un enlace a su repositorio de GitHub.
  - El repositorio debe seguir la plantilla.
  - El LÉAME debería ayudar al lector a navegar por su repositorio. Un buen README ayuda a que su proyecto se destaque de otros proyectos y es el primer archivo que una persona ve cuando se encuentra con su repositorio. Por lo tanto, este archivo debe ser lo suficientemente detallado para enfocarse en su proyecto y cómo lo hace, pero no tanto como para que pierda la atención del lector. Por ejemplo, Proyecto Impresionante tiene una lista seleccionada de archivos README interesantes.
  - Incluya instrucciones breves para replicar completamente el trabajo.
  - La rama del repositorio principal debe mostrar al menos cinco (5) contribuciones sustanciales ciones de cada miembro del equipo.
  - El código tiene que ser:

    Totalmente reproducible.

    Legible e incluir comentarios. En la codificación, como en la escritura, un buen estilo de codificación es fundamental. Te animo a que sigas la guía de estilo de tidyverse.

- Las tablas, figuras y escritos deben ser lo más prolijos posible. Etiquete todas las variables incluidas. Si tiene algo en sus figuras o tablas, espero que se aborden en el texto. Las tablas deben seguir el formato AER.