Universidad de los Andes

Etapa 2

Automatización y uso de modelos de analítica de textos

Grupo 33:

Natalia Ricaurte

Valentina Lara

Manuela Lizcano

Bogotá D.C

1. **Proceso de automatización**

Enfoque de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelos y acceso por medio de API

Para la construcción del proceso de automatización realizamos un seguimiento de la etapa 1, donde se desarrolló en primera instancia el entendimiento de los datos, seguido de esto realizamos el procesamiento de los datos y por último implementamos los modelos: Supported Vector Machine, Regresion Logística, Random Forest, KNN, Multinomial NB y Arboles de decisión. Esto con el objetivo de encontrar el modelo cuyas métricas de F1-score, exactitud, recall y precisión mostraran la mejor ejecución.

Para la implementación de la etapa 2 del proyecto, tomamos la misma lógica explicada anteriormente para lograr construir el pipeline y lograr desplegar este en la API. A continuación, presentamos una descripción del proceso que fue realizado por el ingeniero de datos para el desarrollo de este.

1. **Construcción del pipeline:**

En cuanto a la construcción del pipeline para la automatización de preparación de. Datos y entrenamiento del modelo de clasificación se cargaron y se corrigieron los datos sea un archivo de Excel o CSV, para asegurar que el texto esté en un formato adecuado de gramática y escritura. Luego se utilizó un pipeline que transforma los textos en vectores numéricos mediante TF-IDF. En este se incluye como tokenizer una función encargada de preparar los textos en este caso convertir todo en minúsculas, corregir caracteres, tokenizar el texto en palabras individuales, eliminar las palabras irrelevantes en este caso las stopwords que no aportan gran valor a la clasificación, al igual que la lematización para reducir las palabras y tener la forma base para tener un análisis con mayor exactitud. Estos pasos aseguran que el texto este en aptas condiciones para así poder entrenar el modelo de clasificación en este caso con la utilización de Random Forest, en donde se predijo las categorías de ODS. Los datos se dividieron en entrenamiento y prueba para poder ajustar el modelo y verificar el rendimiento, y al finalizar el entrenamiento el pipeline genera las respectivas predicciones y evalúa las métricas relevantes para clasificación. De igual forma dentro de este mismo, se decidió analizar las palabras más frecuentes para tener un mejor análisis sobre las características de cada ODS. A partir de esto el modelo entrenado se guarda el pipeline.joblib para poder reutilizarlo sin necesidad de volverlo a entrenar y que se automatice el proceso de preparación y predicción del modelo

2. **API:**

En cuanto al proceso de la API se implementó con el objetivo de lograr acceder a las funcionalidades de predicciones y para el entrenamiento del pipeline. Para la implementación del API, utilizamos como framework FastAPI. Este fue usado con el objetivo de lograr construir y gestionar correctamente los endpoints por medio de los principios de REST. De igual forma, el API ayudo a realizar el procesamiento de los datos y análisis de una manera eficiente y rápida. Específicamente relacionado con el Fondo de Poblaciones de las Naciones Unidas, creamos dos endpoints, donde el primero se enfocaba en predecir las opiniones de los ciudadanos y finalmente se encuentra el

endpoint enfocado en el entrenamiento del modelo, usando grandes cantidades de datos.

El primer endpoint, recibe el texto en el formato de JSON, y en primer lugar son procesados por el pipeline para el preprocesamiento y para el modelo de Random Forest . El pipeline primero transforma el texto en vectores numéricos usando TF-IDF y luego esto es pasado por el clasificador de Random Forest. El resultado de este pipeline es la predicción con las probabilidades de pertenencia a cada una de las tres clases. Los resultados de las predicciones son devueltas en la misma secuencia en la se mandaron las opiniones.

El segundo endpoint, está diseñado para recibir un documento con múltiples datos, en el formato de excel, este documento contine tanto las opiniones como la variable objetivo, el cual en nuestro caso es la ODS. El proceso de reentrenamiento es realizado por medio del conjunto de datos obtenidos en los pasos de preprocesamiento anteriores. Después de realizar el reentrenamiento, se guarda una versión del pipeline y despues esta es reemplazada por modelo binaro que fue entrenado anteriormente. El resultado de este endpoint es un conjunto de métricas como F1-score, exactitud, precisión y recall, las cuales permiten identificar el funcionamiento del modelo.

El API logra automatizar el proceso de preprocesamiento de los datos y la prediccion y reentrenamiento de los datos.

El reentrenamiento de los datos podía ser realizado mediante 3 técnicas:

- Utilizando Random Forest, entrenar el modelo juntando los datos antiguos con aquellos nuevos que eran ingresados por los usuarios.
    - Ventajas: Considerando que el reentramiento de los datos se está haciendo en grandes volúmenes de datos, esto puede hacer que el modelo sea más completo y que permita una mejor ejecución durante la clasificación, de igual manera al contar con múltiples opiniones redactadas por varios usuarios estos pueden llegar a eliminar el sobreajuste del modelo hacia ciertos comportamientos.
    - Desventajas: El problema principal de este enfoque se relaciona con que, dado que la cantidad de datos usados para reentrenar el modelo es bastante grande. Esto puede llegar a hacer que el modelo consuma de memoria innecesaria, haciendo que se arruine la ejecución.

- Utilizando Random Forest, entrenar el modelo únicamente usando los datos nuevos ingresados por el usuario.
    - Ventajas: este tipo de implementación permite que el modelo se enfoque completamente en el comentario que se acaba de realizar por el usuario, haciendo que sea un enfoque relevante especialmente si las opiniones antiguas ya no muestran significancia alguna sobre la construcción del modelo actual y de la problemática que se está viviendo. Considerando que los datos que van a ser usados para el entrenamiento son únicamente aquellos ingresados por el usuario, esto va a disminuir radicalmente el tiempo de ejecución considerando que la cantidad de datos que son usados para reentrenar el modelo es menor.
    - Desventajas: este tipo de implementación puede llegar a tener falencias en la predicción, considerando que puede llevar a causar que exista cierta perdida de datos relevantes que estaba usando el modelo para el proceso de la clasificación. De igual forma, es posible que, al considerar únicamente opiniones recientes, esto pueda llevar a causar sesgos en la clasificación, haciendo que el modelo no tenga el mejor rendimiento.

- Utilizando Random Forest, sin embargo, partiendo de cierta cantidad de datos que ya han sido entrenados previamente, junto con los datos nuevos.
    - Ventajas: Considerando que cierta proporción de los datos ya han sido entrenados correctamente, esto hace que el tiempo de ejecución se de en un tiempo relativamente corto, considerando que no es necesario entrenar el modelo desde cero. Adicional a esto, el modelo logra brindar ciertos atributos de clasificación usando la generalización en aquellos datos que ya han sido entrenados anteriormente, y diferenciación con los datos nuevos, para lograr identificar nuevos comportamientos en las opiniones más recientes.
    - Desventaja: Este modelo, tiene una gran desventaja frente a los demás enfoques considerando que, si el modelo inicial está construido con errores o fallas en la precisión, esto mismo se va a ver en los nuevos resultados. De igual forma, esto puede llevar que el modelo se sobreajuste considerando si existen diferencias muy grandes entre los datos anteriormente entrenados y los nuevos.

Considerando estas tres alternativas, tomamos la decisión de entrenar el modelo, únicamente haciendo uso de los datos nuevos. La decisión fue tomada precisamente porque el objetivo de esta aplicación es lograr clasificar opiniones de los ciudadanos para identificar las problemáticas principales relacionadas con los objetivos de desarrollo sostenible, ODS 3, 4 y 5. Considerando que los datos son opiniones personales, estos datos pueden llegar a variar drásticamente considerando la posición política, la cultura y los principios de cada individuo, es por esta razón, necesitamos únicamente tener en cuenta opiniones recientes que contengan información relevante para que el proceso de clasificación tenga un mejor desempeño.

2. Desarrollo de la aplicación

El diseño y construcción de esta aplicación tiene como propósito facilitar el proceso de clasificación de las opiniones de los ciudadanos acerca: Bienestar y Salud, Educación e Igualdad de Género. Esta herramienta optimiza el trabajo que debe ser realizado por los trabajadores internos del Fondo de Poblaciones de las Naciones Unidas, por medio de la reducción de tiempos en la ejecución de la tarea de clasificar las opiniones de los ciudadanos que viven día a día problemas relacionadas con las temáticas mencionadas anteriormente. A partir de esta clasificación, se puede llegar a facilitar la toma de decisiones enfocadas en las problemáticas más relevantes que se estén viviendo en ciertas poblaciones.

Justificación de como funcionamiento de interfaz de usuario

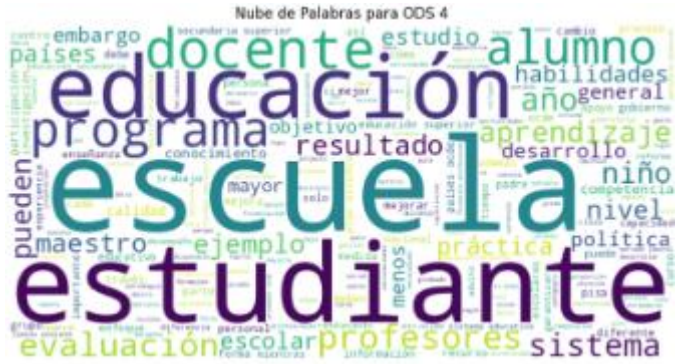
Para concretar un poco el funcionamiento de nuestra interfaz de usuario contamos con dos diferentes secciones: predicción y entrenamiento. Para predecir, iniciamos con una pequeña introducción de cómo funciona el sistema y su funcionalidad principal. De igual forma incluimos una descripción de cada uno de los 3 ODS, con el objetivo de que el usuario logre identificar las posibles opiniones que pueden llegar a ser consideradas por parte del Fondo de Poblaciones de las Naciones Unidas. Seguido de esto, se encuentra el espacio de texto en donde se ingresa la opinión (en caso de quere ingresar mas de una opinion se debe separar por diferentes lineas de texto) y por último se encuentran los datos obtenidos después de que se ejecute el pipeline, junto con los respectivos resultados de las probabilidades de pertenecía a cada una de las tres clases.

En la pestaña de entrenar, recibe un documento de Excel, con múltiples registros de opiniones con el objetivo de obtener los resultados de las métricas a partir de los datos ingresados. De igual forma, incluimos las palabras que fueron usadas para la realización de la clasificación y por último justificamos nuestro enfoque en los datos que fueron usados para el reentramiento.

Resultados probables: Para el caso de que la opinión tenga palabras como

ODS 3

ODS 4



ODS 5

Para todos estos casos hay que tener en cuenta que a pesar de que la probabilidad de pertenencia a una clase puede llegar a ser mayor, de igual forma el modelo obtiene las probabilidades de pertenencia a cada una de las tres clases.

Trabajo en equipo

- Para la elaboración de este proyecto, Natalia Ricaurte fue encargada de cumplir con el rol de líder de equipo, entre las funciones importantes fue establecer las fechas en donde se iban a realizar las reuniones, de igual forma establecimos los pre-entregables del grupo y verificamos que las asignaciones de carga de trabajo fueran realizadas de forma equitativa.
    - Natalia configuro 2 reuniones, en donde en la primera nos encargamos de realizar el lanzamiento y planeación para la elaboración de esta etapa del proyecto, de igual forma, tuvimos las reuniones de seguimiento donde adelantamos la elaboración del proyecto y finalmente realizamos la entrega final del proyecto.

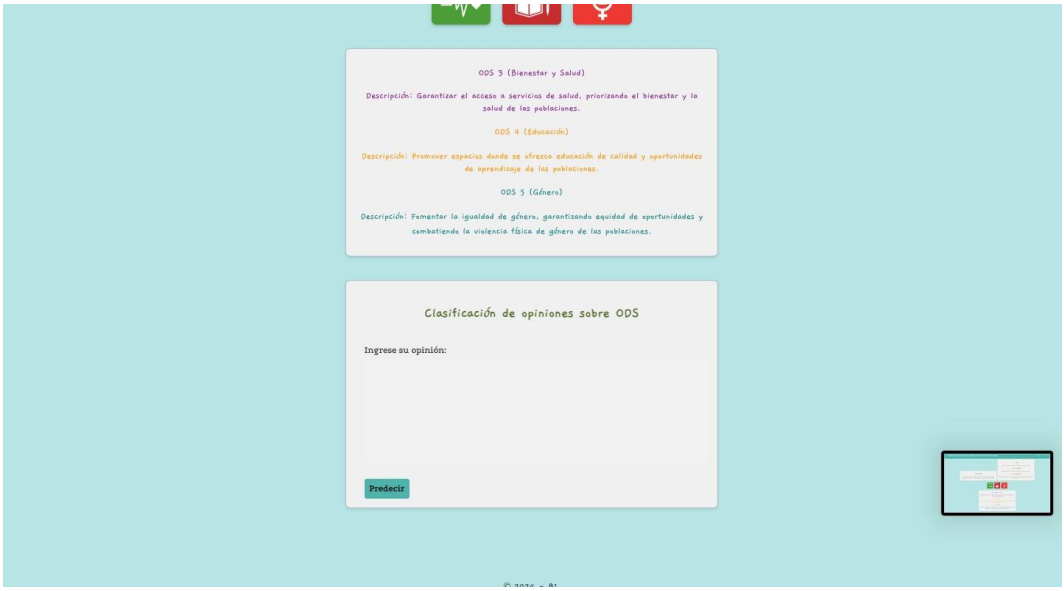
- El ingeniero de datos fue asignado a la labor de Natalia Ricaurte, la cual tenía como labor lograr revisar la calidad del proceso de automatización específicamente relacionada con la construcción de este modelo.
- El ingeniero d software responsable del diseño y de la aplicación y resultado estuvo encargada Valentina Lara, la cual se encargó específicamente de elaborar el diseño de la aplicación y generar el video.
- El ingeniero de software responsable del diseño de la aplicación y resultados fue responsable Manuela Lizcano, específicamente se encargó en liderar el diseño de la aplicación.

- Tiempo dedicado total: 10 horas

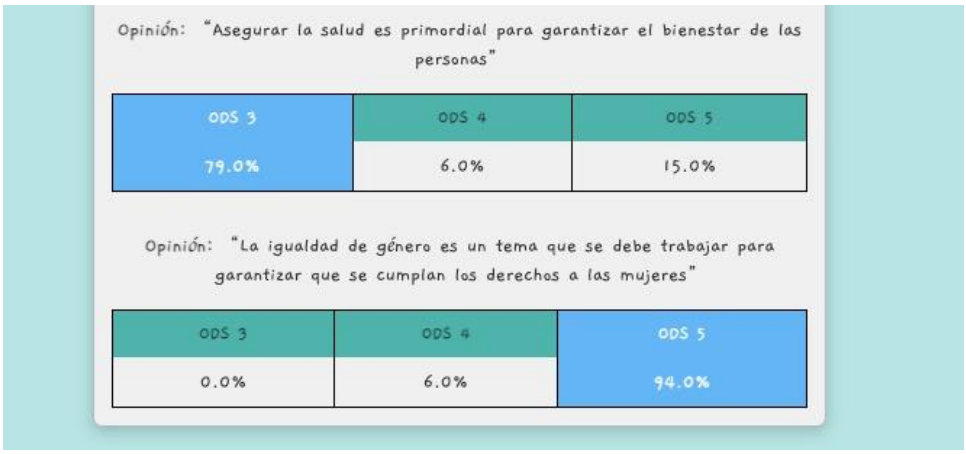
- Retos
    - El reto principal para la elaboración de esta etapa del proyecto fue lograr implementar el pipeline considerando que el manejo de errores que fueron surgiendo en el camino requerían de diferentes cambios. De igual forma, el mantenimiento de las actualizaciones del pipeline hizo que el proceso fuera más largo, especialmente considerando cuando se realizaron los cambios del enunciado.
    - Otro reto que tuvimos que enfrentar fue la conectividad entre el backend y el frontend, considerando que inicialmente el entorno virtual no contaba con ciertas dependencias que eran necesarias para la ejecución de la aplicación. De igual manera, considerando que los ambientes de las maquinas personales de cada integrante del grupo eran diferentes, existían ciertas versiones donde la aplicación funcionaba y otras tocaba cambiarlas, haciendo que el proceso se demorara fuera más demorado.
    - Finalmente, un reto tuvimos que enfrentar estuvo relacionado con que lo lográbamos decidir concretamente que incluir en la aplicación, especialmente considerando que queríamos que el API estuviera construido de la manera más sencilla posible, para que la experiencia de usuario fuera la mejor.

- Planeación vs Realidad
    - Durante la realización de esta etapa del proyecto, existieron varias consideraciones que tuvimos que cambiar cuando estábamos ejecutando la interfaz del usuario. Inicialmente teníamos un diseño definido, el cual contaba con múltiples características para que esta fuera más fácil de usar y que los resultados obtenidos fueran desplegados de cierta manera. Sin embargo, al final optamos por crear una interfaz de usuario mucho más practica y amigable para el usuario, esto con el objetivo de facilitar el proceso.

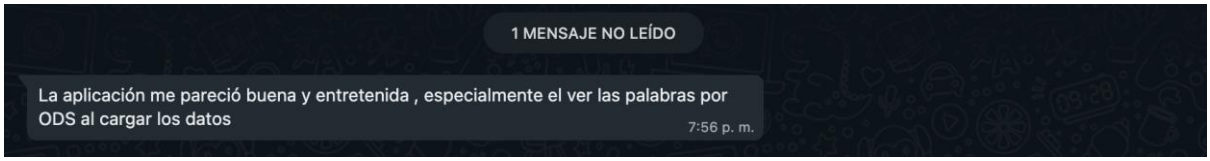
- Validación
    -

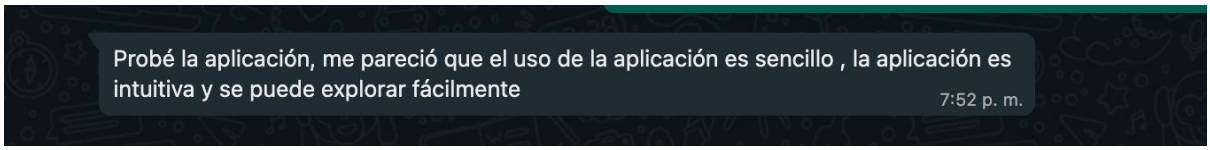
- 
ODS 3 (Bienestar y Salud)
Descripción: Garantizar el acceso a servicios de salud, priorizando el bienestar y la salud de las poblaciones.
ODS 4 (Educación)
Descripción: Promover espacios donde se ofrezca educación de calidad y oportunidades de aprendizaje de las poblaciones.
ODS 5 (Género)
Descripción: Fomentar la igualdad de género, garantizando equidad de oportunidades y combatiendo la violencia física de género de las poblaciones.
Clasificación de opiniones sobre ODS
Ingrese su opinión:
Predecir

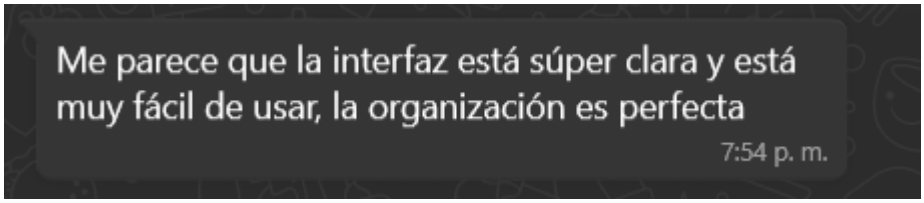

Clasificación de Opiniones Objetivos de Desarrollo Sostenible
Predecir
Entrenar
Bienvenidos
Esta aplicación tiene como objetivo recolectar información a partir de opiniones ciudadanas con el objetivo de tomar decisiones por medio del aprendizaje automático.
Guía
Esta aplicación incluye una sección para predecir y otra para entrenar.
Para predecir:
Ingrese una opinión y obtenga la clasificación de la misma en los ODS.
Para entrenar:
Ingrese un archivo con opiniones y su clasificación en los ODS para entrenar el modelo (archivo xlxs).
ODS 3 (Bienestar y Salud)
Descripción: Garantizar el acceso a servicios de salud, priorizando el bienestar y la salud de las poblaciones.
ODS 4 (Educación)
Descripción: Promover espacios donde se ofrezca educación de calidad y oportunidades de aprendizaje de las poblaciones.
ODS 5 (Género)
Descripción: Fomentar la igualdad de género, garantizando equidad de oportunidades y combatiendo la violencia física de género de las poblaciones.



Comentarios de usuarios que probaron la aplicación:







- Repartición de 100 puntos
    - Valentina Lara: 1/3 de los puntos totales
    - Natalia Ricaurte: 1/3 de los puntos totales
    - Manuela Lizcano: 1/3 de los puntos totales