**Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API:**

**Aclaración:**

Como para la limpieza de los datos se utilizaron funciones personalizadas, al utilizar el pipeline en la aplicación se debían tener las mismas funciones creadas. Por lo anterior, la limpieza de los datos se realizó directamente en el back de la aplicación.

**Construcción del Pipeline:**

El pipeline se compone principalmente por dos pasos:

****

El vectorizador, en este caso se utiliza TF-IDF con n-gramas, y el modelo de clasificación, en este caso RandomForest.

El pipeline es entrenado con los datos de entrenamiento (valga la redundancia) y el modelo es evaluado con las diferentes métricas:

**Texto

Descripción generada automáticamente**

**Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media**

Finalmente, el archivo es exportado usando la librería “joblib”.

Texto

Descripción generada automáticamente

**Construcción de la aplicación:**

Para la construcción de la aplicación podemos destacar dos partes principales:

**Backend:**

Aquí se realiza la creación de la API utilizando “Fastapi”. Primero se realiza la configuración CORS para permitir que cualquier origen haga peticiones al backend. y se carga el pipeline.

Texto

Descripción generada automáticamente

Se descargan las herramientas para la tokenización, stopwords y lematizado.

Texto

Descripción generada automáticamente

Se definen funciones para corregir caracteres, eliminar palabras consecutivas duplicadas, y ajustar textos para mejorar su limpieza y preprocesamiento:

Texto

Descripción generada automáticamente

Estas funciones limpian y preparan los textos para ser ingresados al modelo de clasificación.

Se definen las estructuras que recibe el API, como la entrada de texto para la predicción y los nuevos datos para reentrenamiento:

Texto

Descripción generada automáticamente

Y finalmente se definen los endpoints:

Captura de pantalla de computadora

Descripción generada automáticamente

El endpoint /predict recibe una lista de textos, los preprocesa, y utiliza el pipeline para predecir la clase y la probabilidad.

Texto

Descripción generada automáticamente

El endpoint /retrain permite subir un archivo con nuevos datos para reentrenar el modelo. Los datos se validan, preprocesan, y se vuelven a entrenar usando el pipeline. Este endpoint admite archivos CSV o Excel, combina los nuevos datos con el conjunto de datos existente y vuelve a entrenar el modelo. Por último, las métricas de evaluación se devuelven como respuesta:

Texto

Descripción generada automáticamente

**Frontend:**

Para la interfaz se desarrolló una aplicación básica que ofrece las dos funcionalidades al usuario, clasificar opiniones y reentrenar el modelo:

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Si el usuario elige clasificar una opinión, puede hacerlo escribiendo distintas opiniones en el textbox o subiendo un archivo CSV con opiniones para su clasificación.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Si el usuario desea realizar el reentrenamiento del modelo, es posible hacerlo subiendo un archivo CSV a través de la interfaz

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Desarrollo de la aplicación y justificación**

La idea detrás del desarrollo de esta aplicación es brindar una herramienta útil para automatizar la clasificación de información textual brindada por ciudadanos con los diferentes Objetivos de Desarrollo Sostenible (ODS). Esta es una tarea que consume una gran cantidad de recursos para el Fondo de Poblaciones de las Naciones Unidas, entre ellos, la participación constante de expertos para la clasificación de dicha información.

Identificamos principalmente dos actores que hacen parte de la organización que usarán activamente la aplicación y se beneficiarán de ello.

**Analista de políticas públicas:**

**Conexión con el proceso de negocio:** Este rol utilizará la aplicación para analizar opiniones ciudadanas con relación a los ODS 3, 4 y 5, lo cual impacta directamente en la toma de decisiones para la creación de políticas públicas más informadas. La herramienta apoya el proceso de desarrollo de políticas basadas en datos, que pueden identificar problemas clave y medir el impacto de las políticas en salud, educación y equidad de género.

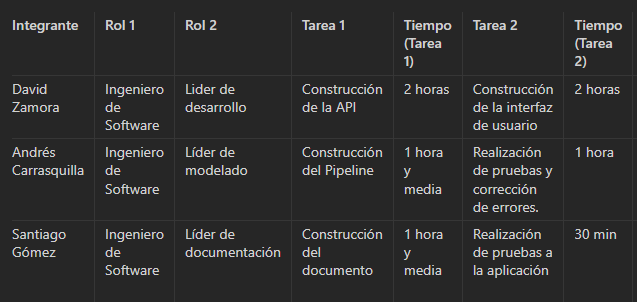
**Importancia:** Para este rol, la existencia de la aplicación es crucial ya que sin ella no podrían tener acceso a un análisis rápido y detallado de las opiniones ciudadanas. La automatización en la clasificación de textos les permite enfocar sus esfuerzos en la creación y ajuste de políticas, en lugar de invertir tiempo en la recopilación y análisis manual de datos.

**Investigador de datos:**

**Conexión con el proceso de negocio:** El investigador de datos es responsable de generar reportes y estudios basados en las opiniones ciudadanas. En este caso, como la aplicación va a ayudar en la tarea de clasificación de las opiniones, va a apoyar directamente el proceso de generación de reportes. Este proceso es fundamental para evaluar el impacto social de las políticas y para identificar problemas clave relacionados con los ODS.

**Importancia:** La aplicación reduce considerablemente el tiempo y esfuerzo requeridos para analizar grandes cantidades de datos, permitiendo a los investigadores centrarse en la interpretación de resultados y la generación de insights. Sin esta herramienta, el proceso de investigación sería mucho más lento y menos eficiente.

**Trabajo en equipo**

****

**Tabla

Descripción generada automáticamente con confianza media**