

Proyecto 1 Etapa 2 – Construcción de modelos de analítica de Textos

INTELIGENCIA DE NEGOCIOS

Juan Andrés Eslava Tovar

Santiago Osorio Osorio

Alejandro Segura

Departamento de Ingeniería de Sistemas y Computación

Universidad de los Andes

12 de octubre de 2024

1. Introducción

En la primera etapa de este proyecto, se construyeron modelos de analítica de textos utilizando técnicas de aprendizaje automático para analizar textos relacionados con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. Estos modelos permitieron identificar patrones en las opiniones de los ciudadanos y proporcionar predicciones sobre el cumplimiento de estos objetivos en Colombia. La metodología incluyó la comparación de varios algoritmos, siendo la Regresión Logística el modelo más efectivo, con un rendimiento promedio del 75% en precisión, recall y F1-score.

Para llevar estos modelos a un entorno operativo y lograr un impacto práctico, es crucial automatizar el proceso de análisis y permitir que los usuarios interactúen con los modelos de manera eficiente y sencilla. En esta Etapa 2, el objetivo principal es la automatización y uso de los modelos de analítica de textos, con un enfoque en crear un sistema que permita el análisis continuo y la interacción amigable para los usuarios.

El sistema se basará en dos tecnologías clave:

- **Front-end desarrollado en React:** Una interfaz de usuario dinámica y responsive que permitirá a los usuarios finales interactuar con el modelo, ingresando texto en lenguaje natural y recibiendo predicciones rápidas sobre el ODS relacionado.
- **Back-end desarrollado con FastAPI:** Una API eficiente y robusta para gestionar las solicitudes y procesar los datos de entrada. FastAPI permitirá que el modelo analítico se integre con la aplicación de manera rápida y segura, soportando tanto la predicción como el reentrenamiento del modelo.

Los objetivos específicos de esta etapa son:

- **Automatizar el proceso de construcción, evaluación y reentrenamiento de los modelos**, asegurando que estos se mantengan actualizados y precisos a lo largo del tiempo.
- **Desarrollar una aplicación web basada en React**, que permita a los usuarios interactuar de forma sencilla con los modelos de análisis de texto y recibir predicciones basadas en datos ingresados.
- **Implementar una API REST utilizando FastAPI**, que permitirá a la aplicación interactuar con los modelos para obtener predicciones y ejecutar procesos de reentrenamiento cuando sea necesario.

La automatización del proceso reducirá significativamente el esfuerzo manual necesario para analizar grandes volúmenes de texto, permitiendo que las organizaciones gubernamentales y ONGs tomen decisiones más rápidas y basadas en datos actualizados. Además, la aplicación web proporcionará una experiencia de

usuario optimizada, facilitando la adopción de esta tecnología por parte de los usuarios no técnicos.

Con la combinación de React y FastAPI, el sistema garantizará una **alta eficiencia**, escalabilidad y facilidad de uso, lo que permitirá su despliegue en diversos escenarios, tanto en organizaciones públicas como privadas, contribuyendo de manera significativa a la medición y monitoreo de los ODS en Colombia.

2. Proceso de Automatización del Modelo

2.1. Descripción general del proceso

El proceso de automatización del modelo tiene como objetivo garantizar que los modelos de analítica de texto se puedan utilizar de manera continua y eficiente, sin la necesidad de intervención manual. Para lograr esto, se implementó un pipeline que cubre las siguientes etapas: preparación de datos, construcción del modelo, evaluación del desempeño, y persistencia del modelo en un repositorio.

Este proceso está diseñado para ser escalable y flexible, permitiendo que el sistema gestione de manera autónoma tanto las predicciones como los reentrenamientos del modelo a lo largo del tiempo. Todo esto es gestionado a través de una API REST desarrollada con FastAPI que actúa como interfaz entre la aplicación y el modelo analítico.

2.2. Pipeline de Automatización

El pipeline automatizado consta de las siguientes fases:

- **Preparación de los datos:**
 - **Recepción de nuevos datos:** Los datos de entrada se reciben en formato JSON a través de la API. Estos datos son transformados para garantizar que sigan el formato requerido por el modelo.
 - **Preprocesamiento de textos:** Se aplican técnicas de limpieza como la eliminación de stopwords, lematización, y transformación a vectores numéricos. Este proceso se realiza automáticamente cada vez que se reciben nuevos datos.
- **Entrenamiento y evaluación del modelo:**
 - **Entrenamiento del modelo:** Cuando se reciben nuevos datos etiquetados, el pipeline activa un proceso de reentrenamiento. El modelo selecciona los datos relevantes y los entrena utilizando el algoritmo de Regresión Logística, como se vio en la etapa 1.
 - **Evaluación del modelo:** Después del reentrenamiento, se evalúa el desempeño del modelo utilizando métricas como precisión, recall y F1-score. Si el desempeño es aceptable, el modelo actualizado se guarda en el repositorio de modelos.

- **Persistencia del modelo:**

- Una vez entrenado y validado, el modelo es persistido automáticamente en una subcarpeta llamada assets, donde se actualiza también el modelo a usar para el endpoint 1 que predice textos.
- Los logs de cada proceso se almacenan para asegurar trazabilidad y facilitar la detección de errores en caso de fallos.

2.3. API REST

La API fue desarrollada en FastAPI, una herramienta de alto rendimiento que facilita la implementación de APIs RESTful. La API cuenta con dos endpoints principales:

- **Endpoint 1: Predicción de ODS para nuevos textos:**

- Este endpoint recibe solicitudes en formato JSON que contienen uno o más textos en lenguaje natural.
- La API preprocesa los textos, los pasa al modelo entrenado, y devuelve las predicciones junto con la probabilidad de cada predicción. El formato de respuesta también es en JSON, manteniendo la coherencia con los estándares de intercambio de datos.
- Este proceso garantiza que los usuarios finales obtengan resultados rápidos y precisos sobre a qué ODS (3, 4, o 5) está asociado el texto proporcionado.

- **Endpoint 2: Reentrenamiento del modelo:**

- Este endpoint permite enviar conjuntos de datos etiquetados adicionales, que incluyen tanto los textos como las etiquetas correspondientes a los ODS.
- La API recibe estos datos, los incorpora en el pipeline de preparación y activa un proceso de reentrenamiento del modelo. Después del reentrenamiento, se devuelven métricas de desempeño, como el Precision, Recall, y F1-score.
- En cada reentrenamiento, se reemplaza el modelo anterior por el nuevo, garantizando que las predicciones futuras se basen en los datos más recientes.

2.4. Reentrenamiento del modelo

Se definieron tres posibles enfoques de reentrenamiento para mantener el modelo actualizado y efectivo con el tiempo. Cada uno tiene sus ventajas y desventajas:

Se definieron tres enfoques diferentes para el reentrenamiento del modelo, cada uno adaptado a situaciones específicas y con ventajas y desventajas particulares.

- **Reentrenamiento completo con todos los datos:**

- **Descripción:** En este escenario, el modelo se reentrena desde cero utilizando todo el conjunto de datos disponible, tanto los datos nuevos como los datos antiguos. Se trata de una regresión completamente nueva que incorpora todo el historial de información en el proceso de reentrenamiento.
- **Ventajas:** Este enfoque asegura que el modelo esté basado en el conjunto de datos más amplio posible, lo que generalmente aumenta la precisión y la estabilidad del modelo, ya que tiene acceso a toda la información disponible.
- **Desventajas:** Es el proceso más costoso en términos de tiempo y recursos, ya que requiere volver a entrenar el modelo desde cero cada vez que se incluyen nuevos datos. Este enfoque puede no ser eficiente si los datos antiguos ya no son representativos de la situación actual.

- **Reentrenamiento con solo los nuevos datos:**

- **Descripción:** Este enfoque se utiliza cuando el conjunto de datos original está desactualizado o muy antiguo, y el objetivo es reentrenar el modelo únicamente con los datos nuevos y recientes. Se asume que los nuevos datos reflejan mejor la situación actual, por lo que no se utilizan los datos históricos en este proceso de reentrenamiento.
- **Ventajas:** Permite que el modelo se actualice rápidamente sin necesidad de procesar grandes volúmenes de datos históricos, lo que reduce significativamente el tiempo de reentrenamiento y los recursos computacionales necesarios.
- **Desventajas:** Al no utilizar los datos antiguos, el modelo puede perder contexto histórico importante, lo que podría reducir su capacidad de generalización. Este enfoque es útil solo cuando los nuevos datos son significativamente más representativos de la realidad actual.

- **Entrenamiento online:**

- **Descripción:** En este escenario, el modelo se ajusta continuamente a medida que recibe nuevos datos. Este tipo de entrenamiento es incremental, es decir, el modelo se actualiza progresivamente con cada nuevo conjunto de datos sin necesidad de un reentrenamiento completo.
- **Ventajas:** El entrenamiento online es eficiente, ya que el modelo se ajusta de manera continua sin necesidad de un reentrenamiento completo. Esto permite que el modelo se mantenga siempre actualizado y relevante con la incorporación de datos recientes en tiempo real.

- **Desventajas:** Existe el riesgo de que el modelo se ajuste demasiado a los datos más recientes y pierda capacidad de generalización si los datos son ruidosos o si los patrones de los datos cambian drásticamente.

Después de evaluar las ventajas y desventajas de cada enfoque, se decidió implementar el reentrenamiento completo con todos los datos. Esta estrategia garantiza que el modelo siempre considere el conjunto más amplio de datos, lo que mejora su estabilidad y precisión en escenarios donde los datos históricos siguen siendo relevantes para las predicciones actuales. Aunque es más costosa en términos de tiempo y recursos, esta estrategia ofrece una mayor robustez para las predicciones relacionadas con los ODS.

3. Desarrollo de la aplicación y justificación

3.1. Descripción del usuario y rol

La aplicación desarrollada en esta etapa está dirigida a dos tipos de usuarios principales:

- **Funcionarios gubernamentales:** Incluyendo el Departamento Nacional de Planeación (DNP), los Ministerios de Salud, Educación y la Consejería Presidencial para la Equidad de la Mujer. Estos usuarios necesitan monitorear el progreso de los Objetivos de Desarrollo Sostenible (ODS) y tomar decisiones informadas sobre la asignación de recursos en áreas como salud, educación y equidad de género.
 - **Rol:** Los funcionarios utilizan la aplicación para obtener predicciones rápidas sobre el estado de los ODS 3, 4, y 5, basadas en opiniones y datos textuales provenientes de múltiples fuentes (informes ciudadanos, encuestas, redes sociales, etc.). Esto les permite identificar áreas donde los avances son limitados y planificar intervenciones más efectivas.
- **Organizaciones No Gubernamentales (ONGs):** Las ONGs que trabajan en los sectores de salud, educación y equidad de género se beneficiarán de esta aplicación para ajustar sus intervenciones y actividades, alineándose con los datos más recientes sobre el progreso de los ODS.
 - **Rol:** Las ONGs usan la aplicación para recibir insights sobre las áreas críticas que necesitan más apoyo. A través de las predicciones, pueden priorizar sus esfuerzos y dirigir sus recursos hacia donde más se necesiten, con base en la información obtenida de los textos.

3.2. Conexión con el proceso de negocio

La aplicación desempeña un papel fundamental en el proceso de toma de decisiones de los usuarios finales, ya que permite:

- **Monitoreo y evaluación de los ODS:** A través de los resultados analíticos obtenidos de los textos ingresados, la aplicación ayuda a los usuarios a monitorear el progreso en temas críticos como la salud, la educación y la equidad de género. Esto mejora la eficiencia en la gestión de políticas públicas y proyectos de intervención social.
- **Optimización de recursos:** Al identificar áreas que necesitan atención prioritaria, las entidades pueden asignar recursos financieros y humanos de manera más eficiente. Las predicciones también permiten reorientar los esfuerzos hacia los ODS con un rendimiento más bajo, optimizando el impacto de las políticas.
- **Toma de decisiones basada en datos:** La aplicación transforma opiniones y datos textuales no estructurados en predicciones claras y estructuradas. Esto reduce la subjetividad en el proceso de toma de decisiones y facilita la implementación de políticas basadas en evidencia.

En resumen, la aplicación simplifica el proceso de análisis de grandes volúmenes de datos textuales, permitiendo a las organizaciones tomar decisiones informadas y asignar recursos de manera más estratégica, lo que aumenta la efectividad de las políticas y proyectos relacionados con los ODS.

3.3. Funcionalidades de la aplicación

La aplicación, desarrollada en **React** y utilizando una API basada en **FastAPI**, cuenta con varias funcionalidades clave que permiten a los usuarios interactuar fácilmente con los modelos analíticos:

- **Ingreso de textos en lenguaje natural:**
 - Los usuarios pueden ingresar uno o varios textos en lenguaje natural (por ejemplo, informes o comentarios ciudadanos) a través de un campo de texto en la aplicación.
 - Esta funcionalidad es esencial para que los usuarios analicen información nueva en tiempo real sin necesidad de transformar manualmente los datos.
- **Predicciones instantáneas:**
 - Una vez que el usuario ingresa los textos, la aplicación envía los datos al backend gestionado por **FastAPI**, donde el modelo analítico procesa los textos y devuelve una predicción sobre el ODS relevante (3, 4, o 5).
 - Además de la predicción, se muestra la **probabilidad asociada a cada predicción**, lo que permite al usuario evaluar la confianza del modelo en cada resultado.
- **Visualización clara de los resultados:**

- Los resultados se presentan de manera clara y visual, destacando a cuál ODS pertenece cada texto analizado. La visualización incluye gráficos y tablas simples que permiten al usuario interpretar los resultados fácilmente, incluso si no tiene conocimientos técnicos avanzados.
- Esto ayuda a los usuarios a tomar decisiones basadas en datos sin perder tiempo interpretando los resultados del análisis.

4. Resultados

- **Predicción por texto:** La imagen a continuación presenta un ejemplo de las predicciones producidas por el modelo, en la que los textos introducidos han sido categorizados en las correspondientes clases de Objetivos de Desarrollo Sostenible (ODS), junto con la probabilidad de cada predicción.

Inteligencia de Negocios
Predicción por Texto
Re-entrenar Modelo

Predicción por Texto


Ingrese el texto que desea clasificar...

Classificar


Resultados de predicciones:

#	Texto	Clase	Probabilidad
1	Reducir la mortalidad materna.	3	0.63
2	Aumentar la cobertura de vacunación infantil.	3	0.46
3	Asegurar que todos los jóvenes tengan acceso a la educación secundaria.	4	0.68
4	Aumentar el número de docentes cualificados.	4	0.55
5	Eliminar todas las formas de violencia contra mujeres y niñas.	5	0.86
6	Promover la participación igualitaria de mujeres en la política.	5	0.78


- **Métricas del modelo:** A continuación, se muestran los indicadores clave de desempeño del modelo, que incluyen precisión, recall y puntaje F1, que demuestran su eficacia al examinar textos y realizar proyecciones vinculadas a los Objetivos de Desarrollo Sostenible.





Estadísticas del Modelo



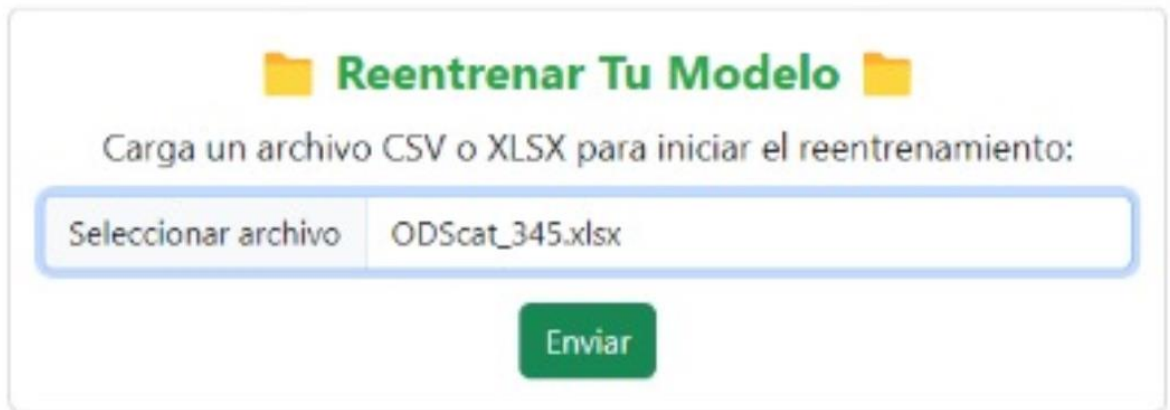
Aquí tienes un resumen de las métricas de rendimiento de tu modelo:


Precisión: 98.08%


Recall: 98.08%


Puntaje F1: 98.08%

- **Re-entrenamiento del Modelo:** Esta imagen presenta la interfaz para llevar a cabo el reentrenamiento del modelo. Los usuarios tienen la posibilidad de cargar un documento CSV o XLSX con información reciente etiquetada para actualizar el modelo.



La interfaz muestra un título "Reentrenar Tu Modelo" con iconos de carpeta a los lados. Debajo, un texto indica "Carga un archivo CSV o XLSX para iniciar el reentrenamiento:". Hay un campo de entrada con un botón "Seleccionar archivo" y el nombre de un archivo "ODScat_345.xlsx". Debajo del campo, hay un botón verde "Enviar".

5. Trabajo en equipo

5.1. Asignación de roles y algoritmos

NOMBRE	ROL
Santiago Osorio Osorio	Lider de proyecto – Ingeniero de software responsable del diseño de la aplicación y resultados.
Juan Andrés Eslava	Ingeniero de datos
Alejandro Segura	Ingeniero de software responsable de desarrollar la aplicación final

5.2. Retos y soluciones

RETOS	SOLUCIONES
Reentrenamiento	Escoger un el reentrenamiento a implementar que diera los resultados deseados.
Desarrollo aplicación	Entendimiento y ruta de aprendizaje para el uso de react y FastApi.

5.3. Reuniones

- Reunión de lanzamiento:
5/10/2024, se entendió el proyecto y se dieron tareas iniciales.
- Reunión de seguimiento de entendimiento:
9/10/2024, se consolidó el entendimiento y preparación, se asignaron los algoritmos.

- Reunión final:

12/10/2024, se consolidó la entrega y se creó el video.

5.4. Reparto de puntos

- Alejandro Segura: 33.3%
- Santiago Osorio: 33.3%
- Juan Eslava: 33.3%