Rodrigo Paz Londoño

Juan José Murillo Aristizábal

Manolo Hernández Rojas

Proyecto Analítica de textos – Etapa 2

En esta etapa, se desarrolló una página web con dos funcionalidades principales; el primero le permite al usuario determinar si una noticia es falsa o no usando el modelo desarrollado en la anterior entrega, y, el segundo, ofrece la posibilidad de reentrenar el modelo y arrojar las métricas actualizadas. En este documento, se detallan varios aspectos relevantes del desarrollo de esta etapa.

1. **Descripción del usuario que va a utilizar la aplicación y sus posibles roles.**

   El usuario que va a utilizar la aplicación es un profesional vinculado a organizaciones que tienen como objetivo monitorear, analizar y contrarrestar la desinformación en el entorno digital, especialmente en el ámbito político. Este usuario puede pertenecer a entidades gubernamentales, medios de comunicación o centros de investigación interesados en la democracia y la opinión pública.

   Entre sus posibles roles se encuentran el de analista de medios y comunicación, responsable de revisar noticias y publicaciones para identificar patrones de desinformación o campañas coordinadas; periodistas o editores, que requieren herramientas de apoyo para verificar la autenticidad de la información antes de publicarla; investigadores, que utilizan este tipo de soluciones para estudiar la propagación y características del lenguaje en las noticias falsas; y asesores de campañas políticas o de imagen pública, que deben responder rápidamente ante la difusión de información engañosa. Asimismo, funcionarios de entidades reguladoras y electorales pueden apoyarse en la aplicación para evaluar el riesgo de desinformación durante periodos clave, como las elecciones, y tomar decisiones oportunas para informar a la ciudadanía.

   Independientemente del rol, la página se convierte en una herramienta clave que permite tomar decisiones más informadas, rápidas y respaldadas por evidencia, mejorando así la capacidad de respuesta ante la desinformación.

2. **Proceso de negocio apoyado y su relación con la página.**

   El proceso de negocio que apoya esta aplicación es el de **verificación y monitoreo de información en medios digitales**, una función crítica dentro de organizaciones encargadas de gestionar la comunicación, proteger la reputación institucional o garantizar la integridad del debate público. Este proceso implica revisar constantemente noticias, publicaciones y contenido en línea para evaluar su veracidad, detectar campañas de desinformación y tomar decisiones estratégicas basadas en esa evaluación.

   La página web desarrollada se integra directamente en este proceso al proporcionar una herramienta automática que permite introducir noticias para que el sistema las analice y determine si se trata de una noticia falsa o verdadera. Además, la aplicación ofrece una predicción acompañada de un nivel de probabilidad. Esto permite que los analistas no solo clasifiquen rápidamente la información, sino que también tengan cierta certeza.

Gracias a esta integración, se reduce significativamente el tiempo y el esfuerzo que normalmente implicaría un proceso manual de verificación, permitiendo así una respuesta más rápida y efectiva ante eventos críticos, como elecciones, crisis políticas o campañas informativas. En última instancia, la aplicación optimiza el proceso de verificación de datos, fortalece la capacidad de análisis institucional y apoya la toma de decisiones basadas en evidencia.

3. **Importancia de la página para roles relaciones verificación y monitoreo de información en medios digitales.**

   La página web desarrollada tiene una importancia fundamental para los roles encargados de la verificación y el monitoreo de información en medios digitales, ya que les permite contar con una herramienta eficiente, automatizada y basada en inteligencia artificial para detectar noticias falsas. Estos roles, que suelen estar a cargo de identificar contenido engañoso, prevenir la propagación de desinformación y proteger la credibilidad de las instituciones, enfrentan diariamente grandes volúmenes de información, lo que hace inviable una revisión manual y exhaustiva de cada pieza de contenido.

   En este contexto, la aplicación se convierte en un aliado estratégico que no solo acelera el proceso de verificación, sino que también mejora su calidad al apoyarse en modelos entrenados con datos reales y criterios objetivos. La posibilidad de obtener una predicción acompañada de una probabilidad le permite al usuario interpretar con mayor claridad los resultados, tomar decisiones informadas y justificar sus acciones ante otros actores dentro de su organización.

4. **Despliegue.**

   - **¿Qué recursos informáticos requiere para entrenar, ejecutar, persistir el modelo analítico y desplegar la aplicación?**

     Entrenamiento del modelo:
     Hardware: Se requiere un equipo con suficiente capacidad de cómputo, con un bueno CPU de alto rendimiento y memoria RAM mínimo de 16 GB, adecuada para procesar grandes volúmenes de datos.

     Software: Entorno Python con librerías especializadas como scikit-learn, spaCy y NLTK y un entorno de virtualización virtualenv.

     Ejecución y persistencia:
     Persistencia del modelo: La API utiliza el formato joblib para guardar el modelo entrenado. Aunque en el código el archivo joblib se carga desde la carpeta models, se cuenta con una carpeta data que almacena el archivo de datos históricos y, que es el csv inicial que se utiliza para crear el modelo.

     Base de datos y archivos: Además de la persistencia del modelo, se requiere un sistema de almacenamiento como SQLite en desarrollo para guardar los datos y configuraciones.

Despliegue de la aplicación:
Servidor de aplicaciones: Uso de Django para la API y el front-end.

- **¿Cómo se integrará la aplicación construida a la organización, estará conectada con algún proceso del negocio o cómo se pondrá a disposición del usuario final?**
  La aplicación se despliega en un servidor local accesible a través de una URL segura, garantizando que solo usuarios autorizados puedan acceder. Se conecta con el proceso de monitoreo y verificación de noticias, permitiendo a analistas y otros roles incorporar rápidamente los resultados del modelo en sus evaluaciones. La API expone endpoints para predicción y reentrenamiento, lo que posibilita su integración con otros sistemas internos como dashboards y sistemas de alertas. La página web, diseñada de manera intuitiva y moderna, facilita el uso tanto de la funcionalidad de predicción como del reentrenamiento, garantizando que los usuarios finales puedan interactuar sin dificultad con la herramienta.

- **¿Qué riesgos tiene para el usuario final usar la aplicación construida?**
  Errores en la predicción:
  Existe la posibilidad de falsos positivos o negativos, lo que podría llevar a decisiones incorrectas si la herramienta se usa de manera aislada sin corroboración adicional.

  Seguridad y privacidad:
  Si no se implementan medidas de seguridad robustas, la aplicación podría ser vulnerable a ataques que comprometan datos sensibles o la integridad del sistema.

  Dependencia tecnológica:
  Una excesiva dependencia en la automatización sin la debida revisión humana puede limitar la capacidad de detectar nuevos patrones de desinformación o adaptarse a cambios en el entorno mediático.

  Mantenimiento y actualización:
  La persistencia del modelo en la carpeta data requiere un manejo adecuado para evitar problemas de inconsistencias que puedan afectar la calidad del reentrenamiento.

5. **Trabajo en equipo.**

   Roles:

   - Líder de Proyecto → Rodrigo Paz Londoño.
     Es quien coordina todas las actividades del equipo, asegurándose de que los entregables se cumplan en los tiempos establecidos y que el trabajo esté bien distribuido entre los integrantes. Define el cronograma interno, organiza las reuniones del grupo, verifica que las tareas asignadas se realicen con calidad, y

actúa como punto de contacto entre el equipo y los docentes o monitores. Además, cuando hay desacuerdos o decisiones clave que tomar, el líder tiene la última palabra.

- Ingeniero de Software y Datos → Juan José Murillo Aristizábal.
  Es el encargado de desarrollar la aplicación web que permite la interacción entre el usuario final y el modelo analítico. Su tarea incluye el diseño e implementación de una interfaz amigable e intuitiva, donde el usuario pueda ingresar textos, recibir predicciones y comprender los resultados de manera clara. También se ocupa de la integración con la API del modelo, la experiencia del usuario (UX), la navegación del sitio y, en algunos casos, la seguridad y el despliegue de la aplicación.

- Ingeniero de Datos→ Manolo Hernández Rojas.
  Este rol tenía la función de diseñar y automatizar los procesos técnicos que permiten preparar los datos, entrenar y mantener el modelo de aprendizaje automático. Su labor incluye la construcción de pipelines para la limpieza, transformación y validación de los datos, así como la implementación de una API que permita al modelo recibir solicitudes y responder con predicciones de forma eficiente. También se encarga de persistir el modelo entrenado y garantizar que pueda ser actualizado con nuevos datos cuando sea necesario.

Horas:

- Rodrigo Paz Londoño: 10.
- Juan José Murillo Aristizábal: 12.
- Manolo Hernández Rojas: 2.

Porcentajes:

- Rodrigo Paz Londoño: 40%.
- Juan José Murillo Aristizábal: 40%.
- Manolo Hernández Rojas: 20%.

6. Desafíos

El principal desafío que presentamos para desarrollar este proyecto fue la elaboración de la API. Esto ocurrió porque no habíamos tenido un acercamiento antes con este tipo de código y estábamos confundidos respecto a cómo se debía implementar. Para solucionarlo, con antelación a la entrega se buscó un framework que fuera fácil de entender y manejar, con la finalidad de que pudiéramos adaptarnos lo más rápido posible. Una vez se tuvo la API, la integración de esta con la página web y demás funcionalidades fueron mucho más simples y rápidas de lograr, haciendo que el restante desarrollo del proyecto fuera mas sencillo y sin muchos obstáculos.

7. Definiciones de Re-entramiento y escogencia del mejor

Re-entrenamiento Incremental: Trata de actualizar el modelo existente incorporando nuevos datos sin empezar desde cero. Se ajusta el modelo ya entrenado de forma progresiva, permitiendo que aprenda de información adicional conforme va llegando. Es eficiente en términos de tiempo y recursos, ya que no requiere volver a procesar toda la base de datos histórica. Si los nuevos datos no están bien balanceados o son muy diferentes de los anteriores, existe el riesgo de que el modelo se desestabilice o acumule errores a lo largo del tiempo.

Re-entrenamiento por Transferencia de Aprendizaje: Consiste en utilizar un modelo preentrenado en una tarea similar y ajustarlo con nuevos datos específicos del problema actual. Se aprovechan las características generales ya aprendidas para luego especializar el modelo en el nuevo dominio. Reduce significativamente el tiempo de entrenamiento y la cantidad de datos necesarios, ya que el modelo ya cuenta con conocimientos previos. Si la tarea de los nuevos datos difiere sustancialmente del original, el modelo podría no adaptarse correctamente y su rendimiento podría verse comprometido.

Re-entrenamiento Completo: Implica volver a entrenar el modelo desde cero utilizando la totalidad de los datos disponibles, es decir, combinando tanto los datos históricos como los nuevos. Esto permite una re-evaluación completa de los patrones y relaciones en la información. Se integra de forma global toda la información disponible, lo que puede conducir a un modelo más robusto y menos sesgado por datos antiguos. Requiere mayor tiempo de procesamiento y recursos computacionales, ya que se vuelve a realizar todo el proceso de entrenamiento.

En esta etapa del proyecto se implementó el re-entrenamiento completo, ya que se combinó la base de datos histórica con los nuevos datos ingresados. Este enfoque permitió actualizar el modelo de forma integral, asegurando que se tenga en cuenta toda la información relevante para mejorar la precisión en la detección de noticias falsas.