

Proyecto 1 – Etapa 2: Analítica de Textos

Universidad de los Andes Departamento de Ingeniería de
Sistemas y Computación ISIS 3301
Inteligencia de Negocios
202320

Santiago Rodriguez Bernal 202011182
Juan Camilo Pacheco Carvajal 201913666

Contenido

1.1	Proceso de automatización del proceso de preparación de datos, construcción del modelo.....	2
1.2	Desarrollo de la aplicación y justificación	2
1.3	Resultados	6
1.4	Trabajo en Equipo	6

1. Automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API

La fase inicial del despliegue de la solución analítica se sumerge de lleno en la automatización integral del proceso, priorizando la preparación exhaustiva de los datos, la construcción del modelo y su persistencia a lo largo del ciclo operativo. Aquí es donde el ingeniero de datos desempeña un papel fundamental, implementando pipelines que orquestan de forma eficiente las tareas de limpieza y transformación de los textos en español. Estas acciones no se limitan a una simple rutina; más bien, se aplican técnicas avanzadas de procesamiento de lenguaje natural que han sido cuidadosamente seleccionadas y ajustadas a lo largo del proyecto, teniendo en cuenta los procesos de construcción de modelos realizados en las etapas iniciales.

Entre las operaciones esenciales que se llevan a cabo durante esta fase se encuentran la eliminación de caracteres especiales, la lematización y la supresión de stopwords. Estas acciones son cruciales para asegurar la calidad de los datos de entrada al modelo, ya que ayudan a normalizar y simplificar el texto, permitiendo que el algoritmo de aprendizaje automático pueda trabajar de manera más efectiva.

En la imagen 1 se puede apreciar el diseño del proceso de automatización en el pipeline. Es importante destacar que el propósito fundamental de este pipeline es automatizar todos y cada uno de los pasos implicados en el procesamiento de los datos, desde su extracción inicial hasta su integración en el modelo de aprendizaje automático. Es por esto que se emplea la biblioteca Scikit Learn, reconocida por su robustez y flexibilidad, especialmente en lo que respecta a la creación de pipelines. La API de Scikit Learn permite la configuración de diferentes pasos de transformación de los datos, los cuales son aplicados secuencialmente, asegurando así un flujo de trabajo coherente y eficiente a lo largo de todo el proceso.

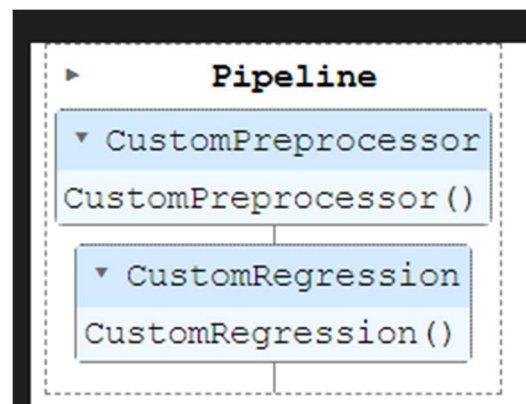


Ilustración 1 Descripción del Pipeline

La construcción del modelo se lleva a cabo mediante técnicas de aprendizaje automático, utilizando algoritmos de clasificación de textos. Durante este proceso, el modelo previamente entrenado es cargado y configurado para interpretar y clasificar textos relacionados con las reseñas de lugares hechas por usuarios y que son de gran importancia para el sector hotelero y turístico del país. Esta etapa es crucial para asegurar que el modelo esté ajustado a las particularidades de los datos y sea capaz de realizar predicciones precisas.

Por otro lado, la persistencia del modelo analítico es esencial para su disponibilidad continua y uso eficiente en el entorno de producción. Se exporta el modelo entrenado en un formato compatible con su reutilización, optimizando su despliegue. La persistencia asegura que el proceso de construcción del modelo no sea necesario en cada interacción, optimizando recursos y acelerando la respuesta del sistema. En este proceso, el pipeline construido donde están encapsulados los diferentes procesos requeridos para la preparación de datos, el entrenamiento del modelo y la predicción como proceso final del mismo, son exportados a un archivo binario o aun archivo “.joblib” utilizando la librería Joblib para posteriormente ser utilizado dentro de la API REST mediante peticiones HTTP.

A continuación, se implementa una API utilizando el framework de Flask. Esta interfaz proporciona un punto de entrada para la interacción con el modelo clasificador. El ingeniero de datos configura la API para recibir solicitudes que contengan textos en español y devuelve predicciones correspondientes a las calificaciones de las reviews. La API actúa como un puente eficaz entre el modelo analítico y los usuarios finales, facilitando la accesibilidad y la integración en diversas aplicaciones.



Ilustración 2 Referencia Flask

Dentro del desarrollo del framework API, Flask destaca por su enfoque eficiente, proporcionando una infraestructura sólida para la interacción entre la aplicación web y el modelo predictivo. La capacidad de Flask para generar automáticamente documentación interactiva es fundamental para el desarrollo de la aplicación, lo que resulta especialmente beneficioso para la colaboración transdisciplinaria. Esto significa que expertos en estadísticas y otros usuarios pueden comprender fácilmente la funcionalidad y los requisitos de la API, incluso sin poseer un profundo conocimiento técnico.

Construimos nuestra API utilizando el framework basado en Python, lo que nos permite dar instrucciones simples para la creación de rutas y la definición de parámetros. Esta simplicidad reduce la complejidad del código y acelera el desarrollo, permitiéndonos centrarnos en la implementación de la lógica empresarial sin tener que lidiar con complicaciones técnicas innecesarias. En resumen, Flask ofrece una combinación ideal de eficiencia y facilidad de uso que hace que sea una elección excelente para desarrollar APIs robustas y escalables.

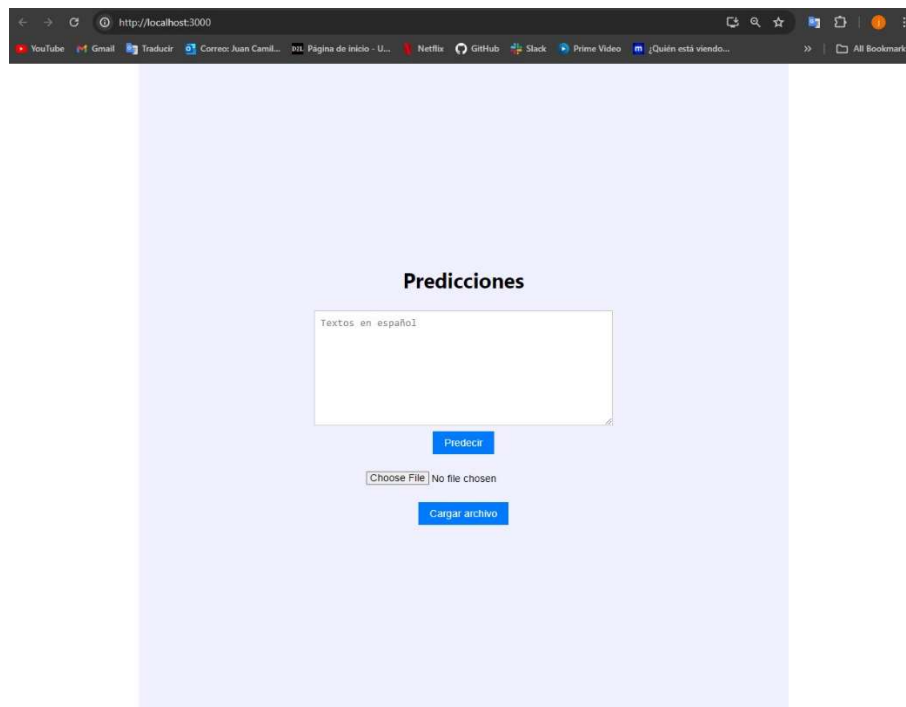


Ilustración 3 Aplicacion desarrollada

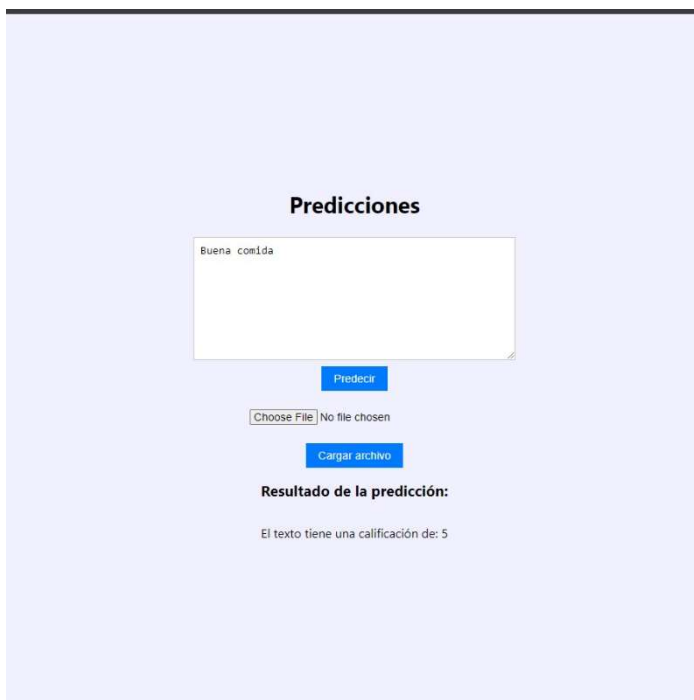
2. Desarrollo de la aplicación y justificación

La aplicación desarrollada se posiciona como una herramienta crucial para los actores estratégicos del análisis turístico en Colombia. Dirigida principalmente a analistas y planificadores del Ministerio de Comercio, Industria y Turismo, así como entidades públicas y colaboradores de la Asociación Hotelera y Turística de Colombia (COTELCO), esta plataforma automatizada se presenta como un recurso esencial para el análisis y la planificación del desarrollo turístico en el país. Al ofrecer un análisis automatizado de las opiniones recopiladas durante procesos de planificación participativa para el desarrollo turístico a nivel territorial, la aplicación busca facilitar la toma de decisiones informadas y estratégicas en el sector.

La conexión entre la aplicación y el proceso de negocio se evidencia en su capacidad para interpretar y clasificar textos según las características clave

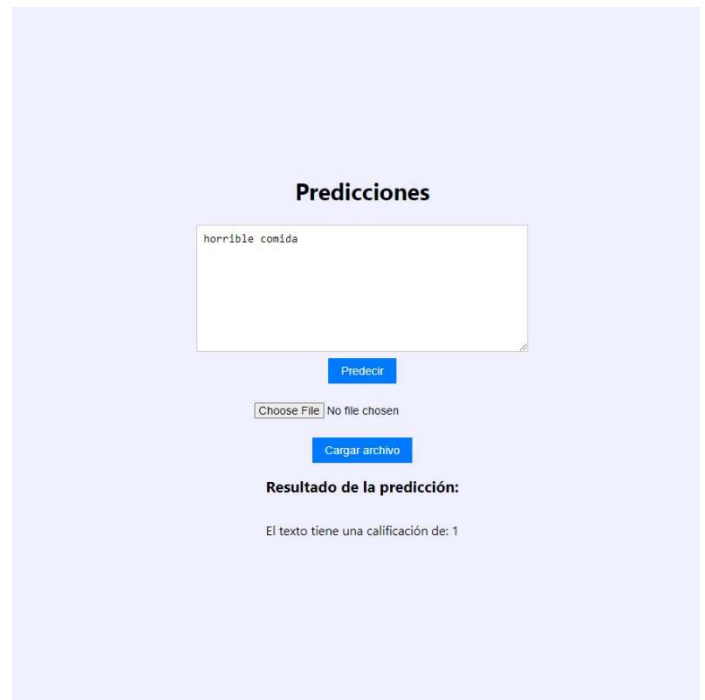
que influyen en la experiencia del turista. Esta funcionalidad se integra perfectamente con las actividades de seguimiento y evaluación de políticas turísticas, permitiendo una identificación más rápida y precisa de problemáticas locales y soluciones efectivas. De esta manera, la aplicación se convierte en un componente estratégico para la toma de decisiones basadas en datos, respaldando la misión de promover el desarrollo sostenible del turismo en Colombia y contribuyendo al diseño e implementación de políticas turísticas efectivas y orientadas a resultados.

Además, el usuario final juega un papel crucial en el aprovechamiento de esta herramienta. Guiado a través de una interfaz amigable, el usuario puede introducir fácilmente textos en español relacionados con cuestiones turísticas, los cuales son automáticamente traducidos en insights significativos y clasificados según los aspectos clave que influyen en la experiencia del turista. Esta capacidad de la aplicación para proporcionar una visión estructurada de las preocupaciones locales empodera a los usuarios con información relevante y detallada para la toma de decisiones estratégicas en el sector turístico. En resumen, la aplicación ofrece una solución integral y eficiente para el análisis y la planificación del desarrollo turístico en Colombia, facilitando la gestión de datos y proporcionando herramientas poderosas para la toma de decisiones informadas.



The screenshot shows a web interface titled "Predicciones". It features a text input field containing "Buena comida". Below the field is a blue "Predecir" button. Underneath the button is a file upload section with a "Choose File" button and the text "No file chosen", followed by a blue "Cargar archivo" button. At the bottom, the text "Resultado de la predicción:" is followed by "El texto tiene una calificación de: 5".

Ilustración 4 Reseña Positiva



The screenshot shows the same web interface titled "Predicciones". The text input field now contains "horrible comida". The "Predecir" button is still present. The file upload section remains the same with "Choose File" and "No file chosen" text, and the "Cargar archivo" button. At the bottom, the text "Resultado de la predicción:" is followed by "El texto tiene una calificación de: 1".

Ilustración 5 Reseña Negativa

Organización/ Apartado	Rol dentro de la empresa	Tipo de Actor	Beneficio	Riesgo
Equipo de Investigación de la Universidad de los Andes	Investigadores	Usuarios Internos	Los investigadores de la Universidad de los Andes emplean el producto de datos para realizar investigaciones académicas centradas en el sector turístico en Colombia. Este enfoque contribuye al avance del conocimiento en el campo del turismo, fortaleciendo la comprensión de los factores que influyen en la industria turística del país y sus implicaciones para el desarrollo económico y social.	Si los datos no son precisos o completos, las investigaciones académicas basadas en ellos podrían carecer de fiabilidad.
Estudiantes de la Universidad de los Andes	Participantes	Usuarios Internos	Los estudiantes pueden aprender sobre los retos y posibles mejoras que tiene el turismo en el país a través de proyectos y análisis basados en el producto de datos, lo que enriquece su educación y genera conciencia sobre las oportunidades dentro del sector	Si la participación de los estudiantes en la recopilación o análisis de datos no se gestiona adecuadamente, podría haber desafíos logísticos o éticos.
Ministerio de Turismo de Colombia	Departamento de Planificación Turística	Cliente	El Ministerio de Turismo podría utilizar los resultados del modelo analítico para tomar decisiones informadas sobre la promoción de destinos turísticos, la asignación de recursos y la elaboración de políticas para impulsar el turismo en Colombia. Esto podría resultar en un aumento del flujo turístico, la generación	Existe el riesgo de que el Ministerio de Turismo no tenga la capacidad suficiente para interpretar y aplicar adecuadamente los resultados del modelo analítico. Esto podría llevar a decisiones inadecuadas o malentendidos en la planificación turística, lo que podría afectar

			de ingresos y el desarrollo económico en las regiones turísticas del país.	negativamente la reputación del Ministerio y el desarrollo del sector turístico en Colombia.
Asociación Hotelera y Turística de Colombia (COTELCO)	Departamento de Desarrollo de Productos Turísticos	Proveedor	COTELCO podría utilizar los resultados del modelo analítico para identificar tendencias en las preferencias de los turistas y desarrollar ofertas turísticas más atractivas y adaptadas a las necesidades del mercado. Esto podría resultar en un aumento de la ocupación hotelera y la rentabilidad para los establecimientos afiliados a la asociación.	Existe el riesgo de que los proveedores no estén dispuestos o no tengan la capacidad de implementar los cambios sugeridos por el modelo analítico. Esto podría limitar el impacto de las mejoras propuestas y llevar a una pérdida de competitividad en el mercado turístico.
Agencia Nacional de Promoción Turística de Colombia (FONTUR)	Departamento de Marketing Turístico	Cliente	FONTUR podría utilizar los resultados del modelo analítico para optimizar sus estrategias de promoción turística, enfocando los recursos en los destinos con mayor potencial de atracción de turistas. Esto podría aumentar la eficacia de las campañas de marketing y contribuir a un mayor posicionamiento de Colombia como destino turístico.	Existe el riesgo de que la implementación de estrategias basadas en los resultados del modelo no genere el impacto esperado, lo que podría llevar a una subutilización de los recursos invertidos en promoción turística y a una disminución de la visibilidad de Colombia como destino turístico.

Empresas de Transporte Turístico en Colombia	Departamento de Operaciones y Logística	Proveedor	Las empresas de transporte turístico podrían utilizar los resultados del modelo analítico para planificar rutas y horarios más eficientes, optimizando la oferta de transporte hacia los destinos turísticos más populares. Esto podría mejorar la experiencia de viaje de los turistas y aumentar la satisfacción del cliente.	Existe el riesgo de que las empresas de transporte no cuenten con la infraestructura necesaria para adaptarse a las demandas generadas por las recomendaciones del modelo analítico. Esto podría resultar en retrasos, cancelaciones o deficiencias en el servicio, lo que afectaría negativamente la experiencia de los turistas y la reputación de las empresas.
Agencias de Viajes en Colombia	Departamento de Ventas y Marketing	Cliente	Las agencias de viajes podrían utilizar los resultados del modelo analítico para personalizar sus ofertas y recomendar destinos y actividades turísticas más relevantes para cada tipo de cliente. Esto podría mejorar la experiencia de reserva de los turistas y aumentar la fidelidad del cliente hacia la agencia.	Existe el riesgo de que las agencias de viajes no cuenten con la capacitación adecuada para interpretar y aplicar los resultados del modelo analítico de manera efectiva. Esto podría llevar a recomendaciones erróneas o malentendidos en la comunicación con los clientes, lo que podría afectar la satisfacción del cliente y la reputación de la agencia.

En conclusión, teniendo en cuenta el mapa de actores anteriormente mostrado esta herramienta no solo beneficia al Ministerio de Comercio, Industria y Turismo y a entidades gubernamentales en la planificación y evaluación de políticas turísticas, sino que también ofrece una oportunidad invaluable para los pequeños hoteles que buscan mejorar su capacidad competitiva en el mercado turístico. Al proporcionar un análisis detallado de las características que influyen en la experiencia del turista y sugerir estrategias de mejora basadas en datos, esta herramienta brinda a los pequeños hoteles la posibilidad de adaptarse y responder de manera efectiva a las demandas cambiantes de los turistas, aumentando así su atractivo y competitividad en el sector.

En un entorno donde la competencia es feroz y la diferenciación es clave, esta herramienta se convierte en un aliado estratégico para los pequeños hoteles, permitiéndoles identificar áreas de mejora y desarrollar estrategias personalizadas para destacarse en el mercado turístico. Además, al ofrecer una visión holística de las preferencias y necesidades de los turistas, esta herramienta ayuda a los pequeños hoteles a mejorar la calidad de sus servicios y experiencias, lo que a su vez puede traducirse en un mayor número de reservas, una mayor satisfacción del cliente y, en última instancia, un aumento en los ingresos.

En última instancia, esta aplicación representa un paso importante hacia el fortalecimiento y la promoción del turismo sostenible en Colombia, beneficiando tanto a los actores gubernamentales como a los empresariales en la búsqueda de un desarrollo turístico más inclusivo y próspero. Al proporcionar a los pequeños hoteles las herramientas y los conocimientos necesarios para competir en el mercado globalizado, esta herramienta contribuye a la creación de un ecosistema turístico más dinámico, diversificado y resiliente, que beneficia a toda la industria y a la economía del país en su conjunto.

3. Interacción con el grupo de Estadística

La definición de la aplicación y su enfoque centrado en el usuario se gestó mediante un proceso colaborativo que destacó por su transdisciplinariedad. En este proceso, ingenieros y estadísticos trabajaron en conjunto para transmitir e integrar sus conocimientos de manera efectiva. Los ingenieros desempeñaron un papel fundamental como desarrolladores de la aplicación, enfocados en satisfacer las necesidades relacionadas con el análisis de datos cualitativos. Por otro lado, los estadísticos participaron desde el rol de usuarios y analistas de datos, probando la aplicación y evaluando su funcionamiento para garantizar su eficacia y cumplimiento de métricas. Su retroalimentación fue crucial para perfeccionar el procesamiento de datos y prepararlos para su posterior análisis.

4. Resultados

Los resultados obtenidos serán mostrados en el vídeo subidos al padlet, donde se muestran las posibilidades que ofrece la aplicación al usuario específicamente la manera en la que se pueden proporcionar nuevos datos escribiendo en un cuadro de texto o subiendo un archivo de testimonios de persona. Estos datos, serán procesados por nuestro modelo y se le mostrará al usuario el resultado de la predicción que obtiene el mismo.

La aplicación final, fue establecida bajo las directrices de una facilidad de uso y una interfaz muy sencilla para el usuario con el fin de que pueda aportar al problema de negocio sin mayores complicaciones. Así mismo, se le otorga al usuario final todo el modelo construido de manera comprimida en el pipeline que se mencionó anteriormente, esto con el fin de darle una persistencia y de qué si es necesario se pueda analizar o replicar todos los pasos de preprocesamiento que se realizaron sobre los datos y cómo esto favoreció en las métricas del modelo.

5. Trabajo en equipo

Estudiantes	Rol	Horas trabajadas	Tareas hechas	Desafíos	Puntaje
Santiago Rodriguez Bernal	Líder de proyecto – Ingeniero de Datos	10	Automatización para el Pipeline – Documento – Vídeo Resultados	Investigación para la automatización	50
Juan Camilo Pacheco Carvajal	Ingeniero de software responsable de diseño y resultados	10	Diseño de la aplicación web – Documento – Vídeo Resultados	Investigación para el uso de API	50

Se realizaron las siguientes reuniones de grupo:

- **Reunión de lanzamiento y planeación:** Se realizó esta reunión el 15 de Abril de 2024. La reunión, con duración de 40 minutos, consistió en definir los roles, la forma de trabajo del grupo y la interacción con el grupo de estadística. Se acuerdan las fechas de reuniones y pre-entregables del grupo, además de los métodos de comunicación para el desarrollo del proyecto.
- **Reunión con el grupo de estadística:** Se realizó esta reunión el 19 de Abril de 2024. La reunión, con duración de 30 minutos, consistió en presentar el problema y solución al grupo de estadística y definir cómo apoyarían este proyecto.
- **Reuniones de seguimiento:** Se realizaron diferentes informes de seguimiento del proyecto mediante mensajes de texto. En estos mensajes se informaba de cada una de las tareas que realizaba cada uno de los integrantes del grupo. Sirven para mantener un registro constante del avance del proyecto y para asegurarse de que las tareas se completen de manera eficiente.