Universidad de los Andes

Facultad de Ingeniería
Departamento de Ingeniería de Sistemas y Computación
Inteligencia de Negocios
2024-10

Proyecto 1: Turismo de los Alpes

Etapa 2: Automatización de Procesos

Grupo 10:
Andrés Arévalo Fajardo - 201923853
María Castro Iregui - 202020850
Mariana Forero Avila - 201922249

6 de abril del 2024
Bogotá D.C.

Tabla de Contenido

Proceso de automatización

El proceso de automatización tuve dos fases. Primero se realizó la automatización del proceso de preparación de datos, construcción del modelo y su persistencia por medio de un *pipeline.* Después, se implementó un API para acceder a las funcionalidades de predicción y de (re) entrenamiento del pipeline. A continuación, se describen las dos etapas en detalle.

Desarrollo del pipeline

El proceso de preparación de datos se realizó mayormente a nivel de la aplicación, por lo que para la automatización del proceso se trataron los datos como datos ya preparados previamente. El pipeline consta con dos etapas principales:

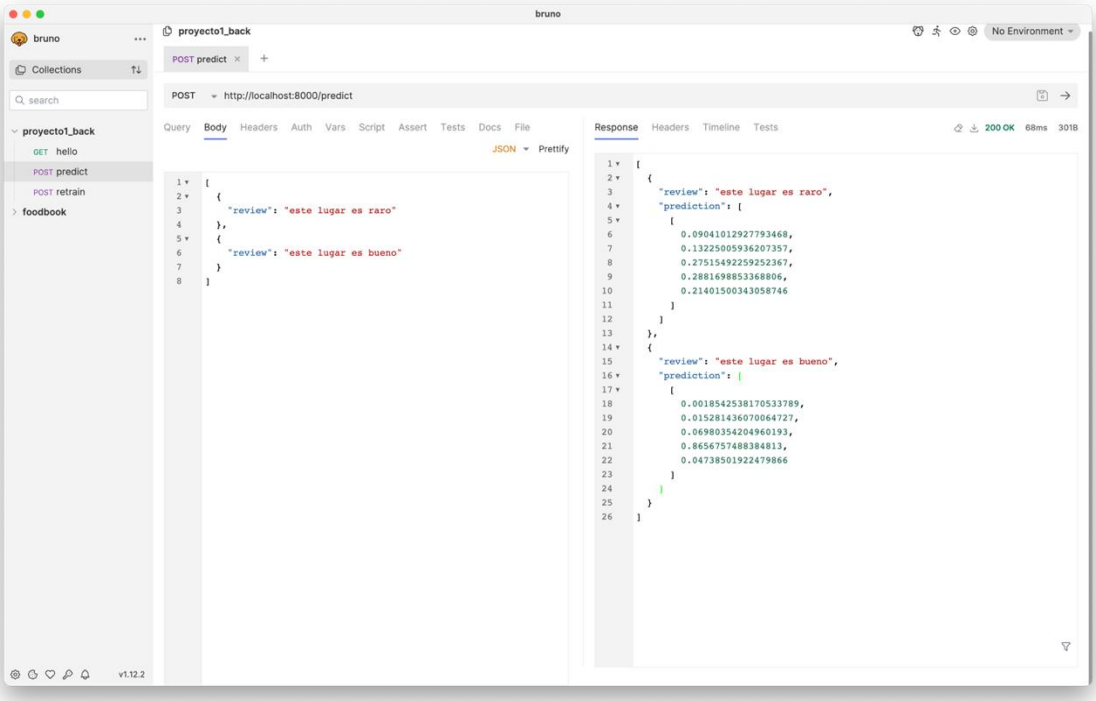
1. La tokenización y vectorización de las reseñas: La vectorización se realiza con el algoritmo de Tf-IDF, al cuál se le pasa el mismo tokenizador realizado para la primera etapa. Este se encarda de eliminar los *stop words,* normaliza el texto para que este todo en minúsculas, elimina los caracteres numericos, la puntiación y los caracteres especiales (no ascii y emojis) y hace use de un lematizador.
2. El clasificador de regresión logística de tipo multinomial y con un máximo de iteraciones de 1000.

Para el entrenamiento de este pipeline de dividen los datos en conjunto de entrenamiento y de prueba, con una división. Luego se hace el ajuste al modelo con los datos de entrenamiento y una validación de las métricas y los top_features. Para esto se definió una función de train_evaluate pipeline. Esta función recibe el pipeline y los datos divididos. Realiza el ajuste y guarda el resultado en formato .joblib usando la función de dump(). Posteriormente, calcula las métricas del modelo y hace uso de otra función get_top_features(), la cual se encarga de devolver las palabras de más alta probabilidad en cada clase. Para la predicción en el pipeline se definió una función que además de hacer la predicción se encarga de realizar el *score* de cada calificación.

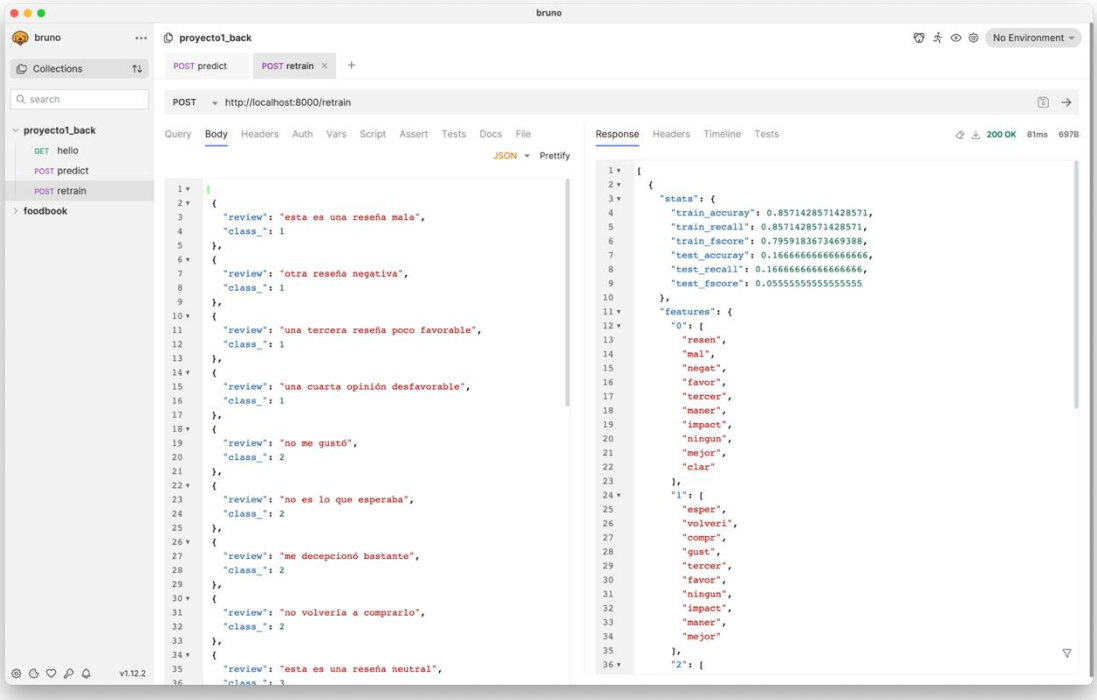
Implementación del API

La implementación del API se realizó con el Framework de FastAPI, este nos permitió exponer los endpoints REST para el acceso al modelo por medio del pipeline previamente construido. En API se encarda de cargar el pipeline desde el formato .joblib de llamar a los métodos correspondientes del pipeline según el endpoint. Para cumplir con las necesidades del negocio, se implementaron dos endpoints, uno para la predicción de reseñas y otro para volver a entrenar el modelo en base a nuevos datos proporcionados por el usuario.

En el caso de la predicción el API recibe en formato JSON un arreglo de reseñas a predecir. La respuesta un nuevamente un arreglo que contiene cada reseña y el *score* que obtuvo para cada clasificación. El *score* más alto corresponde a la clasificación que obtiene la reseña según el modelo. En la siguiente imagen se puede ver el uso del API.



El endpoint para volver a entrenar el modelo recibe en formato JSON los datos. En este caso es un arreglo donde cada elemento tiene la reseña acompañada de su calificación. El API retorna las métricas de *accuracy, recall y* f1-score obtenidas a partir del entrenamiento (tanto del conjunto de entrenamiento com del conjunto de prueba). Además, se incluye por cada clasificación las palabras que más probablidad tienen de corresponder a dicha calificación. A continuación se puede ver el uso del API.



La aplicación web desarrollada se encarga de realizar las transformaciones necesarias para hacer los llamados correctos al API y presentar la información de una manera más digestible para los usuarios.

Desarrollo de la aplicación

Descripción del usuario/rol de la organización que va a utilizar la aplicación

Junto con el grupo de Estadística se determinó que el principal usuario de la aplicación iba a ser el Ministerio de Industria y Comercio de Colombia al cual se le iba a brindar apoyo mediante el desarrollo de un modelo de clasificación basado en técnicas de aprendizaje automático. Esta aplicación tiene el objetivo de hacer que los resultados de este modelo sean más accesibles y fáciles de entender y manejar para que así al misterio pueda utilizarlos para fomentar el turismo en el país.

Aunque esta aplicación se diseñó teniendo al ministerio como principal usuario también puede ser de utilidad para otros actores dentro del contexto como lo son la Asociación Hotelera y Turística de Colombia – COTELCO, cadenas hoteleras de la talla de Hilton, Hoteles Estelar, Holiday Inn y, hoteles pequeños ubicados en diferentes municipios de Colombia. De modo que puedan evaluar los sitios turísticos alrededor de ellos o ver la tendencia positiva o negativa de sus reseñas.

Conexión entre esa aplicación y el proceso de negocio que va a apoyar

La aplicación está vinculada con el proceso de negocio de mejorar y evaluar diferentes sitios turísticos a partir de las reseñas que se tienen de estos, de modo que por un lado se puedan identificar las características que según las reseñas obtenidas hacen que se tenga una percepción positiva o negativa de ese lugar y por otro determinar si un sitio va a ser clasificado con determinada percepción a partir de las reseñas.

Importancia que tiene para ese rol la existencia de esta aplicación

Los beneficios de esta aplicación para el misterio se basan en la eficiencia y efectividad en el análisis de los datos suministrados de modo que se ahorra tiempo y recursos ya que se obtienen resultados de manera casi automática y de forma organizada, contrario a tener que procesarlos manualmente. También está el beneficio de poder tomar decisiones de negocio de manera informada ya que se la facilita al negocio identificar los sitios que están siendo calificados de forma negativa para mejorarlos y a los que se clasifican de forma positiva para recomendarlos. La otra ventaja de la aplicación es que se diseñó pensando en mostrar los

resultados en un formato y forma que hicieran que la información pudiera ser comprendida fácilmente sin tener que ser un experto.

Ajuste tabla de actores

Se realizo la validación de la tabla de actores junto con el grupo de estadística y esta fue aprobada por ellas de modo que este fue el resultado:

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Ministerio de Comercio	Usuario - Cliente	Contribuir al desarrollo y promoción del turismo en Colombia, lo que puede generar beneficios económicos y sociales a largo plazo.	Dependencia excesiva de los resultados del análisis, lo que podría llevar a decisiones erróneas si no se interpreta adecuadamente la información proporcionada.
Cadenas hoteleras y asociaciones turísticas	Financiador	Obtener información detallada y análisis sobre las características de los sitios turísticos, lo que puede ayudar a mejorar la calidad y atractivo de sus destinos turísticos.	Inversión de recursos financieros en un proyecto que podría no cumplir con las expectativas o no generar los resultados esperados.
Universidad de los Andes(entidad colaboradora)	Proveedor	Oportunidad de ofrecer servicios y soluciones especializadas para el análisis de datos turísticos, lo que puede generar ingresos adicionales y mejorar su reputación en el mercado.	Responsabilidad en la calidad y ética de los modelos de clasificación, ya que pueden afectar decisiones importantes en políticas públicas del turismo.
Turistas locales y extranjeros	Beneficiado	Acceso a información más precisa y detallada sobre los sitios turísticos, lo que les permite tomar decisiones informadas y mejorar su experiencia de viaje.	Vulneración de la privacidad y seguridad de los datos personales si la información se maneja de forma inadecuada.

Opciones al momento de definir y desarrollar la aplicación

Desde el principio se tenían claras las dos funcionalidades que iba a tener nuestra aplicación es decir la parte de predicción y de rentrenamiento del modelo. Pero en cuanto a la presentación de los resultados en un principio en la parte de predicciones se tenía pensado solo presentar la calificación con la que había sido clasificada una reseña, pero se pensó que el usuario se podría beneficiar más de ver el porcentaje con el que cada categoría será clasificada en determinado nivel. Ahora después de validar esto con el grupo de estadística nos hicieron darnos cuenta que esta presentación podría no ser muy clara para el cliente. De modo que en vez de utilizar

números para determinar las clasificaciones, las pasamos a categorías que fueran de muy negativa a muy positiva cada una correspondiendo a una calificación en específico. Con esto logramos que los resultados fueran más claros y significativos para el cliente final. Ahora con la parte de predicciones en un principio solo íbamos a mostrar las palabras significativas de cada score como una lista pero hablando con el grupo de estadística nos sugirieron utilizar colores para demostrar la significancia de cada score y poner títulos para cada resultado y no solo el número. También apoyaron nuestra idea de mostrar dos métricas pero con una explicación simple de su significado para que fuera comprensible para el cliente.

En las siguientes imágenes esta parte de la evidencia del trabajo en conjunto y los resultados:

Resultados

Los resultados del proyecto son mostrados en el video adjunto.

Trabajo en Equipo

Roles

- Andrés Arévalo: Ingeniero de datos
- Maria Castro: Líder de proyecto, Ingeniero de software responsable de desarrollar la aplicación final
- Mariana Forero: Ingeniero de software responsable del diseño de la aplicación y resultados

Tiempos

- Reunión de lanzamiento y planeación del trabajo: 1 horas (Todo el grupo)
- Desarrollo Backend: 5 horas (María y Andrés)
- Desarrollo Pipeline y automatización: 4 horas (Andrés)
- Desarrollo Frontend: 4 horas (María)
- Integración y pruebas: 2 horas (Todo el grupo)
- Reunión grupo estadística y validación: 1 hora (Todo el grupo)
- Correcciones aplicación: 2 horas (Todo el grupo)
- Creación del documento: 3 horas (Mariana y María)
- Creación del vídeo: 1 hora (Mariana)

Además de las reuniones, se hizo un seguimiento continuo por medio de canales de texto, por los cuales se informaron de avances, dudas y obstáculos enfrentados.

Retos y Soluciones

Por un lado los retos se centraron en como íbamos a representar las funcionalidades en la aplicación ya que como queríamos que se mostraran los porcentajes con los que cada reseña pertenecía a cada clasificación nos todo investigar un poco más acerca del método de clasificación utilizado pero logramos solucionarlo utilizando ponderaciones. También el tipo de framework que íbamos a utilizar para el front pero nos decidimos por flask. Por otro lado cuadrar nuestros tiempos con el grupo de estadística fue un reto pero al final nos pudimos poner de acuerdo y reunirnos con ellos, también el hecho de que no entendían mucho del lenguaje que manejábamos presento un reto pero con una buena comunicación y paciencia pudimos darnos a entender y sacar resultados en conjunto con ellas.

Repartición Puntos

Andrés Arévalo: 33.33, Maria Castro: 33.33, Mariana Forero: 33.33