Universidad de los Andes

Proyecto 1 Etapa 2

Inteligencia de Negocios

GRUPO PROYECTO 3
Maria Alejandra Estrada García - 202021060
Marilyn Stephany Joven Fonseca - 202021346
Santiago Martínez Novoa - 202112020

Contents

1 Proceso de automatización del proceso de preparación de los datos

1.1 Preparación

Para la automatización del proceso se generaron dos pipelines importantes que fueron exportados para ser usados en la aplicación. En primer lugar, está el pipeline de preprocesamiento, que debido a su configuración y al uso del TF IDF Vectorizer tuvieron que ser exportados separados al modelo. Los procesos que lleva a cabo este pipeline son los siguientes:

- **Limpieza y preprocesamiento general de las reseñas:** En este paso se realiza una serie de operaciones para limpiarlo y prepararlo para su procesamiento adicional. Primero, se eliminan las etiquetas HTML usando expresiones regulares. Luego, se normaliza el texto, eliminando tildes y otros caracteres especiales, y lo convierte todo a minúsculas. Después, identifica y extrae emoticones del texto. Finalmente, elimina caracteres no alfanuméricos y devuelve una lista de palabras limpias.

- **Tokenización:** Se dividió el texto en unidades más pequeñas, o tokens. Posteriormente, se aplicaron métodos para eliminar stopwords, es decir, palabras comunes que no aportan significado, y palabras poco frecuentes que podrían introducir ruido en el análisis.

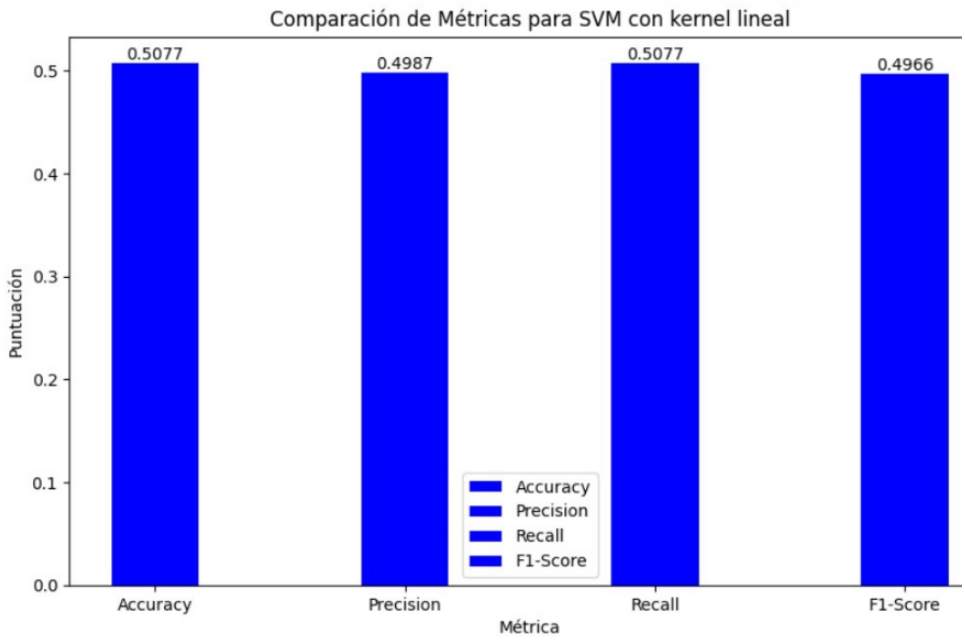
- **Stemming:** Se utilizó en el pipeline el proceso de stemming que es fundamental en el preprocesamiento de texto, ya que ayuda a reducir las palabras a su forma raíz o base.

- **Vectorización:** La vectorización convierte el texto en representaciones numéricas que pueden ser comprendidas por los algoritmos de aprendizaje automático. En este caso, el uso del TF-IDF Vectorizer es especialmente importante, ya que asigna pesos a cada palabra según su frecuencia en el documento y en el corpus en general. Esto ayuda a destacar las palabras más importantes y distintivas de cada documento, algo clave en este caso debido a los objetivos del negocio.

Esta combinación de transformaciones, que incluye el preprocesamiento general, la tokenización, el stemming y la vectorización con TF-IDF, ha demostrado ofrecer las mejores métricas en la etapa 1 de desarrollo. Al simplificar el texto, preservar su significado y resaltar las características distintivas, esta mezcla de transformaciones prepara los datos de manera óptima para el modelo, lo que resulta en un rendimiento mejorado en términos de precisión y generalización.

1.2 Construcción y persistencia del modelo

Para la construcción del modelo se tomó el mejor algoritmo que fue SVM con métricas superiores al 50%, lo que sugiere una clasificación efectiva de los datos. Esto indica que el modelo tiene una alta capacidad para distinguir entre las diferentes calificaciones que se les da a los datos, lo que nos ayuda a obtener una buena predicción. A continuación se muestran las métricas obtenidas:



1.3 Acceso por medio de API

Para poder acceder al modelo creamos un back-end en FastAPI utilizando Python. Se puede acceder a la documentación del API, generada automáticamente por Swagger UI, mediante http://localhost:8000/, una vez se esté ejecutando el backend de acuerdo con las instrucciones dadas en el README del repositorio de github. Adicionalmente se utilizó una base de datos relacional que fue la responsable de que las reseñas anteriormente calificadas se almacenaran y persistieran a través de la diferentes ejecuciones de la aplicación.

Para la implementación de los requerimientos se realizaron los siguientes endpoints para abarcar todas las necesidades del negocio:

Clasificación de reseñas:

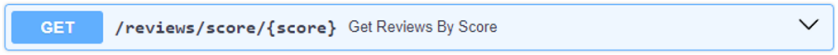
Se definió un endpoint en el que al enviar una reseña individual escrita en tiempo real se pueda obtener una calificación de 1 a 5 teniendo en cuenta el análisis de sentimiento y la predicción del modelo sobre este texto.

Como existe la posibilidad que la información y las reseñas sea demasiadas como para añadirlas individualmente, este endpoint agrega a la base de datos y clasifica múltiples reseñas que se encuentren en un archivo CSV.

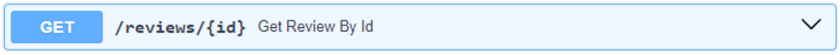
Visualización de reseñas ya calificadas:

Debido a la importancia de visualizar todas las reseñas actuales e históricas, se realizaron los siguientes endpoints.

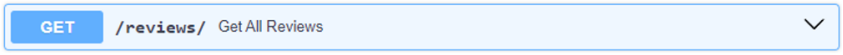
Dado una calificación (score) se retornan todas las reseñas con dicha clasificación.



También es posible visualizar de manera individual cada una de las reseñas.



Por último, para tener una visión de todas las reseñas calificadas y así comprobar la persistencia, se implementó un get all que va almacenando y mostrando todas las reseñas que se han calificado.



Eliminación de reseñas:

Teniendo en cuenta que todas las reseñas se van acumulando, se consideró agregar una opción para borrar todas estas en caso de requerir limpiar el historial. para esto se generaron dos endpoints, el primero permite borrar por completo el historial y el otro permite eliminar una sola reseña específica.

2 Desarrollo de la aplicación y justificación

En primera instancia, el proyecto utiliza una base de datos SQLite para almacenar datos de consultas previas del usuario. SQLAlchemy se utiliza para conectar el backend a esta base de datos. Se guardan los resultados del modelo relevantes para solicitudes futuras, así como los nombres de usuario y contraseñas para la autenticación. Esto evita la necesidad de procesar textos previamente registrados.

El back-end fue desarrollado en FastAPI con Python, este back end permite la comunicación con la base de datos y ofrece los endpoints necesarios, para llevar la operación principal de aplicación.

Por otro lado, el front-end fue llevado a cabo utilizando ReactJS y posee una interfaz intuitiva y confiable que le permite al usuario hacer múltiples acciones:

- Examinar el modelo realizado y conocer tanto sus métricas como los datos que fueron utilizados para su entrenamiento y para su proceso de prueba.
- Acceder al historial de todos los textos que han sido ingresados a la aplicación, ver la clasificación de las reseñas que se les ha otorgado.
- Ingresar textos nuevos a la aplicación escritos manualmente en el input, para que el modelo pueda clasificarlos adecuadamente.

2.1 Rol de la organización que va a utilizar la aplicación y proceso de negocio que va a apoyar

El rol de la organización que utilizará esta aplicación es el de entidades relacionadas con el turismo en Colombia, incluyendo el Ministerio de Comercio, Industria y Turismo, la Asociación Hotelera y Turística de Colombia (COTELCO), así como cadenas hoteleras como Hilton, Hoteles Estelar, Holiday Inn y hoteles más pequeños ubicados en varios municipios del país.

Dado que el objetivo es proveer al sector turístico de Colombia una aplicación que pertmita realizar el analisis cuantitativo de las reseñas recibidas y demostrar que esta puede trabajar con diferentes conjuntos de datos en diferentes momentos, se consideraron los siguientes roles.

Rol	**Tipo de actor**	**Proceso**	**Ventaja**	**Desventaja**
Atención al cliente	Usuario interno	Gestión de Experiencia del Cliente	- Detección temprana de problemas o quejas de los clientes.	- Respuestas inapropiadas o insatisfactorias a los clientes si la clasificación es incorrecta.
Gerencia	Usuario interno	Toma de Decisiones	- Análisis automatizado de reseñas para comprender la percepción del cliente.	- Posibles decisiones erróneas si el modelo de clasificación no es preciso.
Usuarios Externos	Cliente / Usuario	Interacción con la Plataforma de Clasificación de Reseñas	- Acceso a una plataforma fácil de usar para dar reseñas, para además poder confirmar si la calificación dada concuerda con la reseña	- Desconfianza en la plataforma si las clasificaciones no son precisas o consistentes.

Teniendo en cuenta que cada rol tiene un papel importante en el análisis general se consideran claras las acciones para implementar y organizar una a aplicación que pueda brindar los propósitos que cada tipo de usuario puede tener.

2.2 Importancia que tiene para ese rol la existencia de esta aplicación

La existencia de esta aplicación va a aportar una optimicación del proceso de análisis de reseñas. En este un usuario va a poder brindar su experiencia completa y detallada por medio de un texto, y así el resto de actores podrán obtener un análisis automático de dicha reseña, con un

puntaje de lo que el texto general refleja, ya sea que fue muy malo (1), regular (3), muy bueno (5), o una calificación intermedia entre estas.

Con esta comprensión en profundidad de las reseñas, la gerencia puede tomar decisiones estratégicas importantes, como ajustar políticas y procedimientos internos, implementar programas de capacitación para el personal, mejorar la calidad de los servicios ofrecidos, realizar inversiones en infraestructura o tecnología, o incluso cambiar la estrategia de marketing para resaltar los aspectos positivos destacados por los clientes.

Por ejemplo, si la aplicación revela que muchos clientes expresan insatisfacción con la limpieza de las habitaciones, la gerencia puede tomar medidas inmediatas para mejorar los estándares de limpieza y garantizar una experiencia más placentera para los huéspedes. Del mismo modo, si las reseñas resaltan consistentemente la amabilidad y eficiencia del personal, la gerencia puede decidir reforzar estos aspectos positivos a través de incentivos o reconocimientos.

Finalmente, el proceso de negocio que esta aplicación apoyará es el análisis de características de sitios turísticos para comprender qué los hace atractivos para los turistas, ya sean locales o extranjeros. Además, se utilizará para comparar estas características con aquellos sitios que no están recibiendo recomendaciones favorables, lo que afecta negativamente el número de turistas que los visitan. La aplicación también proporcionará un mecanismo para determinar la calificación que un sitio turístico recibe por parte de los turistas, lo que permitirá identificar oportunidades de mejora y estrategias para aumentar la popularidad de los sitios y fomentar el turismo.

3 Resultados

3.1 Mejoras gracias al grupo de estadística

El aporte de Juan Francisco Bernal (j.bernalc@uniandes.edu.co), el estadístico, ha sido esencial para mejorar nuestras habilidades de análisis, lo que nos ha permitido tener una comprensión más completa y precisa de los datos. Gracias a su trabajo, hemos logrado una visión más detallada de los patrones presentes en las reseñas y las calificaciones posibles en un rango de 1 a 5.

A continuación, se destacan las mejoras y sugerencias proporcionadas por el estadístico, que han fortalecido significativamente nuestro enfoque analítico:

- **Revisión y Mejora del Proceso de Preparación de Datos:** La fase de preprocesamiento de datos es crítica para construir modelos precisos. Juan Francisco contribuyó significativamente al revisar y perfeccionar este proceso, asegurando una transformación adecuada de las reseñas.
- **Se consiguió un método estadísticamente significativo:** Luego de realizar las debidas revisiones se determinó que teniendo en cuenta el problema, se consiguió realizar un módelo estadísticamente significativo para entender y predecir los puntajes de las reseñas del sector turístico en Colombia. Lo anterior se puede confirmar dado que se creó una propuesta de valor enfocada en suplir las necesidades del mercado objetivo con una precisión del 51%.

Es importante tener en cuenta que la colaboración con Juan Francisco no solo ha mejorado la calidad y precisión de nuestra aplicación, sino que también ha enriquecido la visualización

de datos y los procesos analíticos. Sus aportes han sido invaluables en el desarrollo de una aplicación más robusta y efectiva para predecir las calificaciones de las reseñas turísticas.

4 Trabajo en equipo

Roles de cada integrante en el equipo y aporte:

Integrantes	**Participación y roles adoptados**
María Alejandra Estrada García	Roles: Líder de proyecto, Ingeniero de software respondable de desarrollar la aplicación final. Acumulado de número de horas dedicadas al proyecto: aproximadamente 20 horas de trabajo neto. Retos enfrentados en el proyecto y formas planteadas para resolverlos: Gestionar eficazmente el tiempo y la Ingeniería de Sistemas y Computación carga de trabajo, implementé una estructura de gestión del tiempo más efectiva y me rodeé de un entorno de trabajo inspirador. Puntos repartidos: 33.33/100.
Santiago Martínez Novoa	Roles: Ingeniero de datos. Acumulado de número de horas dedicadas al proyecto: aproximadamente 20 horas de trabajo neto. Retos enfrentados en el proyecto y formas planteadas para resolverlos: La gestión del tiempo. Para abordar este reto, implementé una planificación detallada, asignando tareas específicas a intervalos regulares y estableciendo plazos realistas. Puntos repartidos: 33.33/100.
Marilyn Stephany Joven Fonseca	Roles: Ingeniero de software responsable del diseño de la aplicación y resultados. Acumulado de número de horas dedicadas al proyecto: aproximadamente 20 horas de trabajo neto. Los retos estuvieron relacionados con la gestión del tiempo fue crucial en este proyecto. Para manejar esta situación, opté por una estrategia de planificación meticulosa. Esta implicó la asignación de tareas específicas en intervalos regulares, así como la fijación de plazos alcanzables y realistas. Puntos repartidos: 33.33/100.

Reuniones de grupo:

- **Reunión de lanzamiento y planeación:** Para definir roles y forma de trabajo del grupo. Se genera lluvia de ideas sobre la forma de resolver el proyecto.

  Día: 13 de abril, 2024 Hora: 3:00pm

- **Reunión de ideación:** Una vez se han explorado los datos del proyecto, la reunión de ideación busca definir la organización/empresa/institución y el rol dentro de ella, que se beneficia de la solución analítica que van a desarrollar

  Día: 15 de abril, 2024 Hora: 5:00pm

- **Reuniones de seguimiento:** Se recomienda mínimo una reunión de seguimiento semanal corta. También pueden ser correos de avance según lo defina el grupo. Pueden tener un tablero de control de las tareas, utilizando herramientas como Trello.

  1. Día: 16 de abril, 2024 Hora: 3:00pm 2. Día: 18 de abril, 2024 Hora: 7:00 am 3. Día: 19 de abril, 2024 Hora: 5:00 pm 4. Día: 20 de abril, 2024 Hora: 10:00 am

- **Reunión de finalización:** Para consolidar el trabajo final, verificar el trabajo del grupo y analizar los puntos a mejorar para la siguiente etapa del proyecto.

  Día: 20 de abril, 2024 Hora: 4:00pm

References

[1] Statista. (2023, 15 octubre). *Colombia: número de hoteles y establecimientos similares 2005-2021.* https://es.statista.com/estadisticas/1018140/evolucion-anual-del-numero-de-hoteles-y-establecimientos-similares-en-colombia/

[2] Elastic. (s. f.). *¿Qué es el procesamiento de lenguaje natural (NLP)? — Una guía completa del NLP.* https://www.elastic.co/es/what-is/natural-language-processing

[3] MATLAB & Simulink. (s. f.). *Support Vector Machine (SVM).* https://la.mathworks.com/discovery/support-vector-machine.html

[4] Fontalvo, W., Mendoza Vega, L., Diaz Solano, B., Hereira, M. J., Torres, D., Diago, F., Martinez, J. (2022). *Caracterización y uso de las redes sociales en las empresas del sector hotelero en la ciudad de Barranquilla - Colombia.* Saber, Ciencia Y Libertad, 17(2), 238–256. https://doi.org/10.18041/2382-3240/saber.2022v17n2.9278

[5] Camilo, Y., & Betancur, F. (2020). *FACTORES DETERMINANTES PARA LA RESERVA DE HOTELES EN LÍNEA EN COLOMBIA. UNA MIRADA DESDE EL MODELO DE ACEPTACIÓN TECNOLÓGICA (TAM) AJUSTADO.* TURPADE. Turismo, Patrimonio Y Desarrollo, 12. https://revistaturpade.lasallebajio.edu.mx/index.php/turpade/article/view/37

[6] DAC. (2019, April 16). *a importancia de la reseñas para los hoteles.* DAC España. DAC España. https://www.dacgroup.com/es/blog/la-importancia-de-la-resenas-para-los-hoteles/

[7] Statista. (2021). *Número de hoteles en Colombia 2021.* https://es.statista.com/estadisticas/1018140/evolucion-anual-del-numero-de-hoteles-y-establecimientos-similares-en-colombia/

[8] Platzi. (s. f.). *¿Como hacer una API con FastAPI?* https://platzi.com/tutoriales/2513-fundamentos-fastapi/13034-como-hacer-una-api-con-fastapi/