Proyecto de analitica de textos

Presentado para la empresa Turismo de los Alpes

Domingo 7 de abril del año 2024

El Ministerio de Comercio, Industria y Turismo de Colombia, 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 están interesados en analizar las características de sitios turísticos que los hacen atractivos para turistas locales o de otros países, ya sea para ir a conocerlos o recomendarlos. De igual manera, quieren comparar las características de dichos sitios, con aquellos que han obtenido bajas recomendaciones y que están afectando el número de turistas que llegan a ellos. Adicionalmente, quieren tener un mecanismo para determinar la calificación que tendrá un sitio por parte de los turistas y así, por ejemplo, aplicar estrategias para identificar oportunidades de mejora que permitan aumentar la popularidad de los sitios y fomentar el turismo.

Para llevar a cabo la realización del proyecto se llevaron a cabo los siguientes pasos a continuación:

a. **(10%) Entendimiento del negocio y enfoque analítico.**

Oportunidad / problema negocio	El Ministerio de Comercio, Industria y Turismo de Colombia, junto con la Asociación Hotelera y Turística de Colombia (COTELCO) y diversas cadenas hoteleras, incluyendo Hilton, Hoteles Estelar y Holiday Inn, así como hoteles más pequeños en diferentes municipios del país, se enfrentan a la necesidad de analizar las características de los sitios turísticos que atraen a los visitantes locales y extranjeros, con el objetivo de comprender qué los hace atractivos o poco recomendables. Además, buscan establecer un mecanismo para evaluar la calidad de los sitios según las reseñas de los turistas, con la esperanza de aumentar el número de turistas que visitan los destinos analizados, mejorar la satisfacción del cliente con la experiencia turística, y posiblemente, incrementar los ingresos generados por el turismo en esas áreas.
Enfoque analítico (Descripción del	Teniendo en cuenta la problemática, el objetivo desde el punto de vista de aprendizaje automático es desarrollar un modelo capaz

requerimiento desde el punto de vista de aprendizaje automático) e incluye las técnicas y algoritmos que propone utilizar.	de clasificar nuevas reseñas en la escala proporcionada (de 1 a 5) donde 5 es que la reseña es muy positiva y 1 es que es muy negativa. Para ello, se propone realizar lo siguiente: **1. Preprocesamiento de datos:** Para trabajar con los algoritmos de aprendizaje de máquina se deberá eliminar los signos de puntuación, caracteres especiales y convertir todo el texto a minúsculas. Luego, se eliminarán palabras vacías (stop words) y se hará una lematización o stemming para reducir las palabras a su forma base. Por último, se hará una vectorización del texto mediante la técnica de TF-IDF para convertir las palabras en vectores numéricos ponderados según su importancia en el contexto de la reseña. **2. Selección de los algoritmos:** Primer Algoritmo: Dado que se trata de un problema de clasificación multiclase con datos textuales, se va a utilizar el de clasificador de Regresión Logística Multinomial. Este algoritmo es adecuado porque es versátil y eficiente, especialmente diseñado para manejar datos textuales y puede manejar múltiples categorías de salida. Además, ya que la técnica TF-IDF permite representar el texto de manera eficiente, capturando la importancia relativa de las palabras en cada documento, permite que esta combinación ofrece un equilibrio entre simplicidad, eficacia y capacidad para manejar datos textuales, lo que lo convierte en una elección sólida para este proyecto. Segundo Algoritmo: Los árboles de decisión son una opción adecuada para clasificar reseñas turísticas debido a su capacidad para aprender relaciones no lineales entre las características del texto y las calificaciones asociadas. Utilizan una estructura jerárquica de reglas de decisión que facilita la interpretación y comprensión de cómo se clasifican las reseñas. Además, son capaces de manejar datos categóricos y numéricos, lo que los hace versátiles para el procesamiento de texto. Tercer Algoritmo: Las máquinas de vector de soporte (SVM) es el tercer algoritmo escogido. Este es un algoritmo de aprendizaje supervisado que se puede utilizar eficazmente para problemas de clasificación multiclase, como lo es este. La idea detrás de las SVM es encontrar el hiperplano que separa mejor las diferentes clases (en este caso, las calificaciones de 1 a 5) en el espacio de características. Este hiperplano se selecciona de manera que la distancia entre él y los puntos de datos más cercanos de cada clase sea maximizada. Esto permite que el modelo tenga un buen desempeño incluso con datos nuevos. **3. Creación del modelo:**

	Para la implementación de cada algoritmo primero se va a dividir los datos en conjuntos de entrenamiento y prueba. Luego, se entrenará el modelo utilizando el conjunto de entrenamiento y se ajustarán los hiper parámetros para mejorar el rendimiento del modelo. **4. Validación del modelo:** Se evaluará el rendimiento de los modelos utilizando métricas adecuadas para la clasificación multiclase, como la precisión, el recall, la puntuación F1 y la matriz de confusión. Después, se elegirá al mejor de los 3 generados y se analizará las métricas para comprender cómo el modelo seleccionado está clasificando las reseñas en las diferentes categorías de la escala. Para el primer enfoque, utilizaremos Regresión Logística Multinomial como algoritmo de clasificación y la selección de palabras representativas mediante TF-IDF permitirá capturar la relevancia semántica de las palabras en cada reseña, lo que mejorará la capacidad del modelo para discernir la intensidad del sentimiento expresado en el texto. En conjunto, estas herramientas proporcionan una solución robusta y eficaz para el objetivo del negocio de analizar y clasificar las reseñas turísticas proporcionando un buen punto de partida para desarrollar el proyecto con el fin de identificar áreas de mejora y promover el turismo. Para el segundo enfoque, como algoritmo de clasificación se eligió Árboles de Decisión, para esto se va hacer un preprocesamiento de los datos para limpar el texto y poder crear el modelo. A partir de este algoritmo, el modelo recibe una entrada de texto ya limpia y transformada y por medio de un árbol de clasificación, clasifica la reseña, o sea le asigna una puntuación entre el 1 y el 5. A partir del resultado se analizarán las métricas de precisión, exactitud, F1, entre otras, además de la matriz de confusión para concluir la calidad del modelo. Para el tercer enfoque, para validar el desempeño del modelo SVM se seguirá un conjunto de pasos, donde el primero consiste inicialmente en evaluar las métricas de clasificación como: la precisión, recall, puntaje F1 y matriz de confusión. Luego en el segundo paso, se analizaron los resultados visualmente por medio de una matriz de confusión en un heatmap. Y finalmente se hará un análisis de errores, revisando la matriz de confusión y los casos donde el modelo hizo predicciones incorrectas, para identificar patrones o características de las reseñas que puedan causar problema de clasificación.
Organización y rol dentro de ella que se beneficia con la	Para el alcance de este proyecto se decide enmarcar como principal beneficiario del proyecto a COTELCO (Asociación Hotelera y Turística de Colombia). COTELCO desempeña un papel fundamental como representante y defensor de los intereses de la

oportunidad definida	industria hotelera y turística en Colombia. Para este proyecto, COTELCO actúa como un facilitador y colaborador clave al proporcionar datos y conocimientos sobre el sector turístico colombiano, así como al participar en la implementación y validación del modelo analítico desarrollado. Además, COTELCO puede utilizar los resultados del proyecto para ofrecer recomendaciones y guías a sus miembros sobre cómo mejorar la calidad de sus servicios y la experiencia del turista, lo que a su vez puede beneficiar a toda la industria turística del país.
Contacto con experto externo al proyecto y detalles de la planeación	A nivel de planeación se concretaron tres reuniones con el equipo de consulta de estadística con Karol Clavijo y Julio Gutiérrez, estas reuniones se harán de manera virtual por medio de google meet para validar el enfoque que le está dando el proyecto. - Fecha primera reunión: Domingo 31 de Marzo - Fecha segunda reunión: Jueves 04 de Abril - Fecha tercera reunión: Sábado 06 de Abril

b. (40%) Entendimiento y preparación de los datos, modelado y evaluación.

En lo que respecta a la fase de entendimiento, preparación de datos, creación de modelo y evaluación de este mismo, se recomienda acceder al siguiente link que lo redireccionará al repositorio de GitHub donde se encuentra de manera detallada el proceso realizado para cada una de estas etapas.

En general, para el entendimiento de los datos se compartieron datos de 7875 reseñas de turistas de los cuales tenemos una única columna object "Review" en el que se encuentra toda la información de la reseña escrita. Por otro lado, hay una única columna numérica "Class" que guarda el nivel de satisfacción del turista con base en la reseña. Para la preparación de los datos de forma general lo que se hizo fue una limpieza de datos, tokenización y vectorización, en específico se hicieron los siguientes cambios:

- Eliminar signos de puntuación, caracteres especiales, tildes y convertir todo a minúsculas.
- Elimina las palabras vacías.
- Lematizar las palabras.
- Convertir números en su representación textual en español en una reseña.
- Vectorizar las palabras de las review.

Para el modelado se plantearon los 3 diferentes algoritmos y se entrenaron con los datos previamente preparados. Para finalizar, en la evaluación se evaluó el f1 score, precisión y recall de cada uno de los modelos para determinar el mejor modelo.

c. (20%) Descripción de los resultados obtenidos

En lo que respecta a los modelos desarrollados se encuentran los siguientes valores de calidad para cada modelo:

Integrante	Modelo	Precision	Recall	F1-Score
Jefferson Hernandez	Regresión logística multinomial	0.47	0.47	0.46
Samuel Goncalves	Arbole de decisión	0.34	0.34	0.34
Ronal Pardo	Máquinas de vector de soporte (SVM)	0.48	0.49	0.48

Para la Regresión Logística Multinomial, el modelo presenta una precisión, recall y F1-Score de alrededor del 0.47. Esto indica que el modelo es capaz de clasificar correctamente alrededor del 47% de las reseñas en la categoría correcta. Aunque esta métrica no es muy alta, aún proporciona un nivel de rendimiento aceptable para el propósito del negocio.

Por otro lado, el modelo basado en Árboles de Decisión muestra una precisión, recall y F1-Score de aproximadamente 0.34. Esto indica que el modelo tiene un rendimiento inferior en comparación con la Regresión Logística Multinomial. Sin embargo, sigue siendo capaz de clasificar las reseñas con una precisión modesta.

Finalmente, el modelo basado en Máquinas de Vector de Soporte (SVM) muestra una precisión, recall y F1-Score de alrededor del 0.48. Este modelo tiene un rendimiento similar al de la Regresión Logística Multinomial, lo que lo convierte en una opción viable para la clasificación de reseñas turísticas por tener un poco

mejor las metricas que el de Regresión Logística Multinomial. Se elige este como el modelo para usar para el proyecto de turismo, el cual tuvo en cuenta las siguiente palabras importantes para clasificar a las reseñas:

Las 30 palabras más positivas: ['bien' 'gente' 'subir' 'gran' 'bueno' 'buena' 'historia' 'agradable' 'viejo' 'desayuno' 'precio' 'siquiera' 'mojitos' 'vendedores' 'dentro' 'demasiado' 'control' 'salon' 'mojito' 'ser' 'habitacion' 'piso' 'caro' 'dos' 'guia' 'bastante' 'visita' 'interesante' 'igual' 'parecio']

Las 30 palabras más negativas: ['peor' 'horrible' 'asco' 'pesima' 'cover' 'pesimo' 'ayer' 'recorrido' 'chico' 'jamas' 'robo' 'rota' 'nunca' 'cubano' 'nardo' 'persona' 'sucias' 'vuelta' 'tardo' 'noma' 'llegada' 'sabado' 'respuesta' 'cero' 'sucia' 'asqueroso' 'quemas' 'lamentable' 'terminar' 'seguridad']

En general, ninguno de los modelos logra alcanzar niveles de precisión, recall o F1-Score que se consideren satisfactorios para una aplicación práctica robusta como el proyecto de Turismo de los Alpes. Sin embargo, mediante la aplicación de técnicas más avanzadas de preparación de datos en el ámbito del procesamiento del lenguaje natural, existe un gran potencial para mejorar el rendimiento de los modelos de clasificación de reseñas turísticas. Estas mejoras podrían conducir a una mejor comprensión y clasificación de las opiniones de los turistas, lo que a su vez podría facilitar la identificación de áreas de mejora en los destinos turísticos y el desarrollo de estrategias más efectivas para promover el turismo.

d. (10%) Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Turistas	Cliente	Se benefician al recibir recomendaciones de viaje más personalizadas basadas en las predicciones del modelo.	Existe el riesgo de que las predicciones del modelo no se alineen con sus preferencias personales, lo que podría llevar a una experiencia de viaje insatisfactoria.
Inversores	Financiador	Pueden ver un retorno de la inversión si el modelo mejora la satisfacción del cliente y aumenta las ventas.	El riesgo es que si el modelo no funciona como se esperaba, puede no ver un retorno de su inversión.

Hoteles, restaurantes, operadores turísticos	Proveedor	Se benefician al recibir más clientes de la empresa de turismo debido a las recomendaciones personalizadas.	El riesgo es que si las predicciones del modelo son inexactas, pueden recibir críticas negativas de los clientes insatisfechos.
Comunidad local	Beneficiados	Se benefician del aumento del turismo y del gasto turístico.	Sin embargo, el aumento del turismo también puede llevar a problemas como el sobre turismo y el desgaste de los recursos locales.

e. (8%) Trabajo en equipo

Integrante	Rol	Tareas realizadas	Horas dedicadas	Retos	Puntos de mejora	Puntaje
Jefferson Hernandez 202120242	Líder de negocio	**Algoritmo trabajado: Regresión logística multinomial** Entendimiento de los datos. Evaluación y análisis de los resultados. Creación del modelo de regresión logística multinomial.	12 hrs	**Encontrar información para el modelo:** Al ser un modelo no visto en clase fue complicado encontrar información para entender y poder aplicar el algoritmo para el problema específico del negocio.	Como mejora hubiera sido mejor estudiar más algoritmos de clasificación para tener un ambiente más amplio y buscar mejores modelos.	**33%**
Samuel Goncalves 202122595	Líder de datos Lider de analitica	**Algoritmo trabajado: Árboles de Decisión** Preparación de los datos	12 hrs	**Transformación y limpieza de los datos:** Fue un reto porque tuve que investigar mucho sobre cómo	Para mejorar a futuro sería mejor estudiar otros modelos y probar diferentes transformaciones para elegir la que mejor se	**33%**

		Modelado de los datos Creación del modelo de Árboles de Decisión		transformar los datos para entrenar los modelos, hubieron muchos retos como entender la lematización, el stem, la vectorización, entre otros.	adapte a lo requerido por el negocio.	
Ronald Pardo 202111309	Líder de marketing Lider de analitica	**Algoritmo trabajado: Máquinas de vector de soporte (SVM)** Desarrollo del documento final Desarrollo de la presentación hecha en Canva para la realización del video Creación de repositorio de GitHub	12 hrs	**Elección del algoritmo**: de tantas posibilidades para escoger, elegir el algoritmo a implementar fue una tarea un poco compleja inicialmente debido a la necesidad de hacer research respecto a posibles algoritmos que funciones para modelos de multi clasificación, como lo implica este proyecto. Sin embargo, gracias a una búsqueda específica y detallada acompañada de preguntas técnicas para apoyo solamente hechas a varias AI y la búsqueda de diferentes fuentes me permitió	Sería ideal tener una mejor organización de tiempo, pero la dificultad de ello implicó que ciertos días se realizará más trabajo que en otros días.	**33%**

				escoger un algoritmo que suele ser utilizado para este tipo de problemas.		