Universidad de los Andes
Inteligencia de Negocios

Proyecto 1: Analítica de textos para reseñas de turismo de los Alpes

Etapa 1 – Construcción de modelos de analítica de texto

Grupo 22
Laura Daniela Arias – 202020621
Nicolás Londoño – 201821364
Daniel Alfredo Reales – 20

7 de abril de 2024

TABLA DE CONTENIDO

CONTEXTO……2
ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO……...3
ENTENDIMIENTO DE LOS DATOS……4
PREPARACIÓN DE LOS DATOS……4
MODELADO Y EVALUACIÓN……5
RESULTADOS……7
MAPA DE ACTORES RELACIONADO CON EL PRODUCTO DE DATOS CREADO……7
TRABAJO EN EQUIPO……8

CONTEXTO

En este proyecto se busca apoyar al 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 que están interesados en analizar las características de sitios turísticos que los hacen atractivos para turistas locales o de otros países y a partir de ese análisis tomar acciones adicionales para poder aplicar estrategias. La tarea se realizará por medio del análisis de información textual proporcionada por estos actores en la forma de reseñas de diversos sitios turísticos. En esta primera etapa se centrará en la construcción de varios modelos de analítica de textos aplicados al caso.

ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

Oportunidad/Problema Negocio	La oportunidad de negocio que se plantea en este contexto es desarrollar un modelo predictivo que permita a los actores del sector turismo (Ministerio de Comercio, Industria y Turismo, COTELCO, cadenas hoteleras) identificar las características clave que hacen que un sitio turístico sea atractivo o no para los visitantes, tanto locales como internacionales. Esto les permitirá: – Comparar las características de los sitios turísticos que tienen altas y bajas recomendaciones, con el fin de identificar oportunidades de mejora. – Determinar la calificación que los turistas otorgan a los sitios turísticos, para poder aplicar estrategias que aumenten su popularidad y fomenten el turismo.

Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar	Dado que los datos proporcionados son, más relevantemente, de naturaleza textual y están acompañados de una categorización numérica, concluimos implementar modelos de clasificación supervisada. De esta manera, podemos buscar predecir la calificación que se le va a otorgar a un sitio a partir de los comentarios realizados. Es así como se decidió que cada estudiante implementara un modelo de los 3 siguientes: – Support Vector Machines o SVM – Naive Bayes – Random Forest Acompañados de un *pipeline* para el manejo de los datos.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	*Los principales beneficiarios de esta oportunidad serían:* El Ministerio de Comercio, Industria y Turismo de Colombia: Podrá utilizar los insights obtenidos para diseñar e implementar políticas y programas que fomenten el desarrollo del turismo en el país. La Asociación Hotelera y Turística de Colombia (COTELCO): Podrá ayudar a sus miembros (cadenas hoteleras y hoteles pequeños) a mejorar sus sitios turísticos y atraer más visitantes. Las cadenas hoteleras (Hilton, Hoteles Estelar, Holiday Inn): Podrán utilizar la información para mejorar la experiencia de los turistas en sus establecimientos y en los sitios turísticos que les interesan. *Dentro de la organización, los roles principales que se beneficiarán serán:* Científicos de datos: Encargados de analizar los datos, construir los modelos predictivos y generar los insights. Ingenieros de datos: Responsables de procesar y preparar los datos, y de desplegar los modelos en un entorno de producción. Ingenieros de software: Encargados de desarrollar las aplicaciones web o móviles

	que permitan a los usuarios finales (turistas, empresas del sector) acceder a los resultados del análisis.
Contacto con experto externo al proyecto y detalles de la planeación	Teniendo en cuenta en número de grupo en la hoja de cálculo se cuenta con el apoyo de dos estudiantes de estadística 1. Estas dos personas son: – Isabella Nova, i.nova@uniandes.edu.co – Ana Sánchez, correo de contacto no proporcionado

ENTENDIMIENTO DE LOS DATOS

Estructura
El archivo completo original cuenta con 7875 datos de reseñas—es decir, 7875 filas—, donde cada entrada cuenta con su reseña textual y su clase de tipo numérico con valores que van del 1 al 5—2 columnas.

Completitud
No se cuenta con dates faltantes dentro de la muestra total. Ningún dato es nulo.

Unicidad
De acuerdo con la información proporcionada por el *Pandas Profiling Report* dentro de los datos se encontraron 29 filas duplicadas, lo cual representa el 0.4% de los datos.

Validez
A pesar de las pocas filas repetidas, todos los datos proporcionados solo válidos para el ejercicio que se ha de realizar.

PREPARACIÓN DE LOS DATOS

Antes que nada, dado que en el entendimiento de los datos encontramos que los datos cuentan con filas repetidas, realizamos una pequeña limpieza donde se busca eliminar todos los datos duplicados. Sin embargo, se tomó la decisión de hacer esta limpieza aún más minuciosa y enfocarla específicamente en repetidos teniendo en cuenta únicamente los valores dentro de la columna de *Review*. De esta manera nos aseguramos de que no contaremos con textos iguales con calificaciones muy distintas que puedan afectar negativamente el modelo. Una vez realizada esta limpieza quedamos con 7802 datos para trabajar.

Dentro de la preparación de los datos se llevaron a cabo diversas acciones para obtener los textos más óptimos para el desarrollo del trabajo. Para un entendimiento más fácil definimos estas acciones en dos grandes categorías—eliminaciones y conversiones—presentadas a continuación:

Eliminaciones

- Caracteres no-ASCII para evitar problemas con elementos no pertenecientes al alfabeto implementado en el español.
- Puntuación que no aporta a la interpretación de la información durante el análisis de cada palabra.
- Palabras vacías o comunes que no aportan al significado o análisis de los textos (e.g. la, el).
- Prefijos y sufijos para así normalizar el texto.

Conversiones

- Conversión a minúsculas de todo el texto para evitar problemas de interpretación por parte de los modelos de palabras iguales con capitalización diferente.
- Conversión de números a su redacción textual para que se acomoden mejor al contexto que se está manejando.
- Lematización de los verbos para reducir la variación de palabra y facilitar el procesamiento e interpretación de los textos.
- Tokenización con el fin de diferenciar mejor las palabras y ofrecer un mejor análisis sobre estas.

MODELADO Y EVALUACIÓN

Como se mencionó anteriormente los 3 modelos que implementamos son SVM, Naive Bayes y Random Forest.

Modelo SVM

El objetivo de un modelo SVM es encontrar un hiperplano en un espacio N-dimensional (donde N es el número de características) que mejor separe las clases de datos. Este hiperplano se elige de manera que maximice la distancia entre los puntos más cercanos de las clases, que se llaman vectores de soporte. De ahí su nombre, *Support Vector Machine*.

Reporte de Entrenamiento para Support Vector Machines

	precision	recall	f1-score	support
1	1.00	1.00	1.00	609
2	1.00	1.00	1.00	936
3	0.98	0.99	0.99	1261
4	0.97	0.97	0.97	1559
5	0.98	0.98	0.98	1876
accuracy			0.98	6241
macro avg	0.99	0.99	0.99	6241
weighted avg	0.98	0.98	0.98	6241

Ilustración 1. Reporte de Entrenamiento para SVM

La anterior imagen ilustra el reporte del entrenamiento del modelo, que presenta métricas bastante favorables. Así mismo, este modelo cuenta con un elevado nivel de precisión de 0.9831757731132831.

Modelo Naive Bayes

Un modelo Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes y en la suposición "ingenua" de independencia condicional entre las características. En el aprendizaje, se calculan las probabilidades a priori de cada clase y las probabilidades condicionales de cada característica dada cada clase en el conjunto de entrenamiento. Para predecir la clase de un nuevo ejemplo, se calcula la probabilidad posterior de cada clase utilizando el teorema de Bayes y la suposición de independencia condicional, y se selecciona la clase con la mayor probabilidad posterior.

Reporte de Entrenamiento para Naive Bayes

	precision	recall	f1-score	support
1	0.95	0.63	0.76	609
2	0.74	0.73	0.73	936
3	0.74	0.67	0.70	1261
4	0.68	0.73	0.70	1559
5	0.76	0.85	0.81	1876
accuracy			0.74	6241
macro avg	0.77	0.72	0.74	6241
weighted avg	0.75	0.74	0.74	6241

Ilustración 2. Reporte de Entrenamiento para NB

La anterior imagen ilustra el reporte del entrenamiento del modelo, el cual cuenta con valores menores a los del modelo anterior pero que no dejan de ser positivos. La precisión para este modelo es de 0.7449126742509213.

Modelo Random Forest

Por último, tenemos el modelo Random Forest. Este modelo funciona combinando un conjunto de árboles de decisión para realizar una predicción. Cada árbol en el bosque se entrena de forma independiente con una muestra aleatoria del conjunto de entrenamiento y una selección aleatoria de características. Para realizar una predicción de regresión, cada árbol en el bosque emite una predicción individual, y la predicción final del bosque se obtiene promediando las predicciones de los árboles individuales.

```
Métricas de rendimiento en el conjunto de entrenamiento:
Error cuadrático medio - Mean Squared Error (MSE): 0.12973781445281204
Error absoluto medio - Mean Absolute Error (MAE): 0.27849863803877584
R cuadrado - R-squared (R^2): 0.9252914771449623
```

Ilustración 3. Métricas de Rendimiento para RF

La anterior imagen ilustra las métricas de rendimiento del modelo. Se observa como este cuenta con un buen ajuste (MSE y MAE bajos) y que puede explicar la variabilidad de los datos de manera efectiva (R^2 cercano a 1).

RESULTADOS

Luego de realizar todo el proceso anterior y construir los 3 modelos propuestos para atender la situación expuesta, llegamos a los siguientes resultados:

Contamos con 3 modelos de clasificación de texto bastante favorables. Todos estos, al ser construidos a partir del mismo conjunto de datos y el mismo procesamiento de los mismos, cuentan con la misma base. Ergo, podemos realizar una recomendación de modelo informada a las entidades relacionadas. Contando con una precisión del 98%, recomendamos el uso del modelo SVM construido para alcanzar los objetivos propuestos.

MAPA DE ACTORES RELACIONADO CON EL PRODUCTO DE DATOS CREADO

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Ministerio de Comercio, Industria y Turismo de Colombia	Cliente/Financiador	Diseñar e implementar políticas y programas más efectivos para fomentar el turismo en el país. Tomar decisiones informadas sobre inversiones en infraestructura y promoción turística.	Posible resistencia al cambio por parte de los actores del sector turismo tradicionales. Dificultad para implementar las recomendaciones a nivel nacional.
Asociación Hotelera y Turística de Colombia (COTELCO)	Beneficiado	Poder brindar a sus miembros (hoteles) información valiosa para mejorar la atracción de sus sitios turísticos. Aumentar el número de turistas que visitan los destinos representados por COTELCO.	Dependencia de la calidad y exactitud de los datos proporcionados por los hoteles miembros. Resistencia de algunos hoteles a compartir información sensible.
Cadenas hoteleras (Hilton, Hoteles Estelar, Holiday Inn)	Beneficiado	Mejorar la experiencia de los turistas en sus establecimientos y en los sitios turísticos de su interés. Aumentar la ocupación y los ingresos de sus hoteles.	Posible recelo a compartir información que pueda beneficiar a la competencia. Dificultad para implementar cambios en los sitios turísticos

			que no les pertenecen.
Turistas (locales e internacionales)	Beneficiado	Tener acceso a información más precisa sobre la calidad y atractivo de los sitios turísticos. Poder planificar viajes más satisfactorios.	Posible desconfianza en la imparcialidad de las evaluaciones de los sitios turísticos. Riesgo de que las mejoras en los sitios turísticos aumenten los precios de los servicios.

TRABAJO EN EQUIPO

Líder de proyecto: Laura
Líder de negocio: Nicolás
Líder de datos: Nicolás
Líder de analítica: Daniel

Reuniones

- Reunión de lanzamiento y planeación
  - Fecha: Marzo 16.
  - Se revisó el enunciado y se definió la dinámica de trabajo.
- Reunión de ideación
  - Fecha: Marzo 17.
  - Definición de los modelos a realizar y quién realiza qué tarea.
- Reuniones de seguimiento
  - De por si no se realizaron reuniones de seguimiento formales, sino que se mantuvo una comunicación constante por medio de un grupo de whatsapp y el compartir de los archivos e información requerida.
- Reunión de finalización:
  - Fecha: Abril 7.
  - Reunión generada a raíz de buscar terminar la entrega en conjunto de la mejor manera posible.