Universidad de los Andes

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

Proyecto 1 Etapa 1

Analítica de Texto

Grupo 28
Gabriel Felipe Dicelis Ramos - 201920847
Juan Pablo Martínez Pineda - 202012623
Laura María Restrepo Palomino - 202013289

Abril 7, 2024
Bogotá D.C.

Tabla de contenido

Introducción

Actores del sector turístico en Colombia, México y Cuba están interesados en analizar las características de sitios turísticos que los hacen o no atractivos para turistas locales o de otros países. En adición, desean contar con un mecanismo que permita determinar la calificación que obtendrá un sitio por parte de los turistas y, de esta manera, identificar oportunidades de mejora que permitan aumentar la popularidad de los sitios y fomentar el turismo.

Para ello, han preparado un conjunto de datos con reseñas de sitios turísticos, donde cada reseña tiene una calificación según el sentimiento que tuvo el turista al visitarlo.

Dado el objetivo de los actores, así como los datos proporcionados, se entiende que se debe realizar una tarea de aprendizaje automático, donde el producto final debe ser un modelo capaz de predecir la calificación de un sitio turístico

Entendimiento del negocio y enfoque analítico

Para esta tarea, es necesario entender el negocio y enfoque analítico, para proporcionar contexto y objetivos claros en torno a la solución que se busca proponer.

Oportunidad/problema negocio	Al contar con datos históricos relacionados a las reseñas dejadas por turistas, junto con una calificación numérica que permite clasificar el nivel de satisfacción, se puede identificar qué aspectos contribuyen a que un sitio turístico sea o no atractivo para los turistas. Así, se puede contribuir a la mejora de los lugares turísticos, o evitar construir establecimientos cercanos a aquellos sitios con malas calificaciones.
Enfoque analítico	El enfoque analítico se hará con respecto a palabras clave dentro de las reseñas que tengan más peso la calificación que le dará un usuario. Esto permite conocer específicamente qué aspectos a mejorar.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Las cadenas hoteleras se pueden beneficiar de esta oportunidad de varias maneras: ● Aquellos hoteles que ya se encuentran cerca a sitios turísticos pueden crear convenios para ayudar a su mejora y atraer más huéspedes. ● Se pueden construir hoteles cercanos a los mejores sitios turísticos y así atraer más huéspedes. ● Se pueden ofrecer eventos turísticos desde el hotel teniendo en cuenta aquellos sitios con mejores características. Esto ofrece una ruta alternativa de ingresos para los hoteles. ● Habiendo atraído más clientes, se pueden generar más empleos.

Contacto con experto externo al proyecto y detalles de la planeación	• Barak Valderrama - b.valderramaa@uniandes.edu.co • Laura Valendia - la.valendial@uniandes.edu.co Fecha de reunión: 5 de abril del 2024, 5:30 p.m. Canal: Zoom

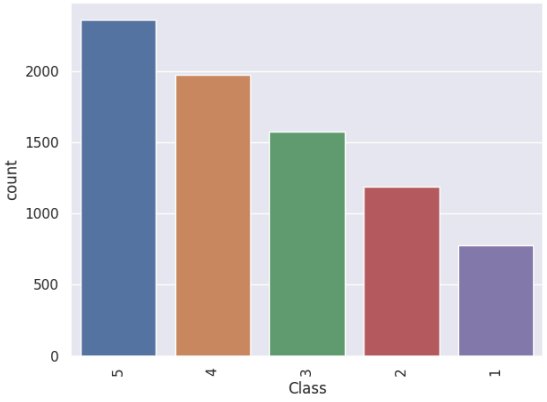
Entendimiento y preparación de los datos

Entendimiento de los datos

Habiendo establecido los objetivos claramente, se pasa a hacer una exploración de los datos proporcionados. Estos datos cuentan con 7875 registros separados en dos valores: reseñas y calificación (respectivamente etiquetados como *Review* y *Class*). Las reseñas son una variable de tipo no estructurado, puesto que no cuentan con valor numérico y no se pueden clasificar nominal u ordinalmente. Por otro lado, la calificación es una variable de tipo categórica, pues se encuentra en un rango definido (1 al 5, números enteros).

Al hacer un análisis más detallado de las reseñas, se observa que el promedio de caracteres por reseña es aproximadamente 408, mientras que el promedio de palabras por reseña es aproximadamente 71. También se tiene que alrededor del 99.92% de las reseñas se encuentran en español, mientras que el resto se dividen en inglés, italiano, portugués y albanés.[1]

Por parte de la calificación, se observa que el valor con mayor conteo es 5 (alrededor de 2500 incidencias), seguido por 4, 3, 2 y 1.



Cuenta de clasificación

Verificación de calidad

Después de haber analizado los datos en su estado puro, el siguiente paso es verificar la calidad de ellos. Para esto, se explora su completitud, unicidad y validez.

[1] Para detectar el idioma se hizo uso de la librería langdetect.

Por parte de la completitud, se verificó que los valores de la variables no fueran nulos y se obtuvo que ningún registro contaba con valores vacíos. Por parte de la unicidad, se verificó que no hubiera datos duplicados y se obtuvo no fue cierto pues se encontraron 102 datos repetidos, lo cual no puede ocurrir, pues contar con los mismos datos múltiples veces puede afectar el modelo de entrenamiento. Por último, en cuanto a la validez, se verificó que la variable de clasificación solo contara con valores enteros del 1 al 5 y se obtuvo que esto era cierto al no haber ningún valor atípico.

Es importante notar que no se hizo un análisis de consistencia, pues según la información proporcionada no hay variables que deban ser consistentes entre sí.

Limpieza y preparación

Una vez se han entendido los datos con los que se va a trabajar, se procede a hacer una limpieza y preparación de ellos, de tal manera que el modelo pueda ser correctamente entrenado. La limpieza contó con las siguientes tareas:

- Eliminación de los registros duplicados, para evitar darles más peso en el modelo.
- Eliminación de los registros que no estuvieran en español, en aras de normalizar el idioma de las reseñas.
- Eliminación de caracteres que no fueran ASCII, para simplificar el texto a estudiar.
- Conversión de las palabras a minúsculas, para estandarizar el texto.
- Eliminación de puntuación, porque no aporta al modelo al no ser palabras.
- Reemplazo de valores numéricos a su representación textual, nuevamente para normalización.
- Eliminación de *stopwords* o palabras comunes (para reducir el número de palabras que no aportan significado).

Esta limpieza termina eliminando alrededor de 0.01% de los registros.

Comentarios de los expertos:
“En la medida de lo posible, no eliminar datos, como se está haciendo con los que están en otro idioma”.
Respuesta:
“Esto solo se hizo puesto que la cantidad de reseñas en otro idioma era alrededor de 6 (0.0008%), un valor insignificante.”

Por parte de la preparación, primero se tokenizaron las palabras, es decir, se separaron individualmente. Esto es útil para el siguiente paso: normalización. Aquí se convierte cada palabra en su forma base (ejemplo: verbos como “corriendo”,

“corrió”, “correrá” se cambian a “correr”), permitiendo estandarizar las reseñas y tener un análisis más preciso.[2]

Finalmente, se selecciona la calificación como variable objetivo, pues es la que se busca predecir, usando una técnica de frecuencia de término y frecuencia inversa de documento. A partir de esto, se crean los conjuntos de entrenamiento y prueba, destinando el 80% de los datos para el entrenamiento.

Modelado y evaluación

Con un conjunto de datos limpios y preparados, se procede a crear tres modelos que ayuden en la tarea de predecir la calificación de una reseña. Vale la pena mencionar que como se está trabajando con una variable categórica, esta es una tarea de clasificación.

Naive Bayes - Gabriel Dicelis

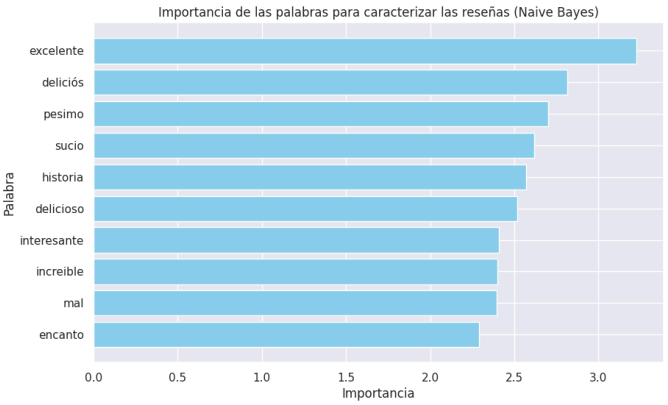
El algoritmo de Naive Bayes utiliza el concepto de independencia condicional utilizado en estadística para calcular la probabilidad de que una instancia (como una reseña) pertenezca a una determinada clase (como "positiva" o "negativa") dada la presencia de ciertas características (como palabras específicas).

En el caso de las reseñas, cada palabra en la reseña se trata como un "bloque" individual de información. Naive Bayes asume que la presencia de una palabra en la reseña no está influenciada por la presencia de otras palabras en la misma reseña. Por ejemplo, si una reseña contiene la palabra "excelente", Naive Bayes asume que la aparición de esta palabra es independiente de la presencia de otras palabras en la misma reseña.

En resumen, el algoritmo de Naive Bayes trata cada palabra en una reseña como una entidad independiente y calcula la probabilidad de que la reseña pertenezca a una determinada clase utilizando el teorema de Bayes y la suposición de independencia condicional entre las palabras.

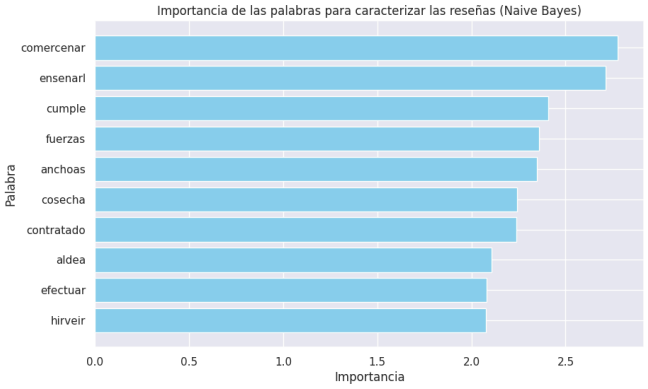
Como resultado del entrenamiento del modelo, se obtuvieron estos resultados con una precisión del 41%

[2] En los anexos se puede encontrar un archivo .csv con el nuevo conjunto de datos limpio y preparado.



Importancia palabras iteración 1 NB

Se observa que la mayoría de las palabras son adjetivos, para el objetivo de negocio que es encontrar elementos diferenciadores en las reseñas, estos resultados no son muy significativos, por lo que se realiza una segunda iteración sin adjetivos.



Importancia palabras iteración 2 NB

Regresión Logística - Juan Pablo Martínez

La regresión logística es un tipo de modelo estadístico (también conocido como modelo logit) que se utiliza a menudo para clasificación y análisis predictivo. Este modelo es un clasificador de aprendizaje automático supervisado que extrae las características de los valores reales que son dados como entrada, multiplica cada una por un peso, las suma y pasa esta suma a través de una función sigmoide para generar una probabilidad, por lo tanto, se produce un valor entre 0 y 1. En el caso de clasificación binaria, si la probabilidad obtenida es inferior a 0.5, se retorna 0. Por otro lado, si la probabilidad obtenida es superior a 0.5, se retorna 1.

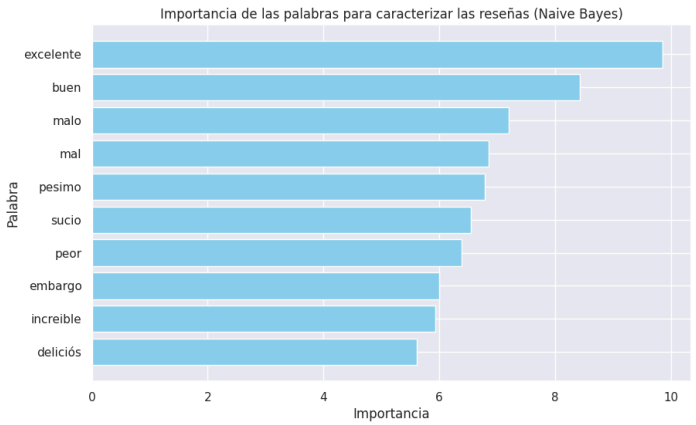
Similar a la regresión lineal, la regresión logística también se utiliza para estimar la relación entre una variable dependiente y una o más variables independientes, pero se utiliza para hacer una predicción sobre una variable categórica.

La regresión logística también es utilizada para clasificación multinomial, es decir, de varias clases. La variable dependiente tiene tres o más resultados posibles; sin

embargo, estos valores no tienen un orden especificado. En este caso, en lugar de utilizar una función sigmoide para obtener el valor de la probabilidad, se usa una función *softmax*. Dado un número N de clases, la función obtiene N valores de probabilidad (uno por clase) y clasifica esa instancia en la clase que haya obtenido el mayor valor.

Al igual que en otros modelos de aprendizaje automático, se busca minimizar la función de pérdida (loss) para obtener el modelo que mejor se ajuste a los datos y obtener el estimador de máxima verosimilitud. Esto se realiza utilizando el proceso de descenso de gradiente estocástico, el cual es un algoritmo de optimización iterativo que permite encontrar los mínimos locales de una función, en el que la idea es tomar pasos de manera repetida en dirección contraria al gradiente de la función.

Después de entrenar el modelo con los datos proporcionados, las métricas obtenidas para *accuracy*, *precision* y *recall* tuvieron un valor aproximado de 0,47. Por otro lado, al calcular el F1 score se obtuvo un valor de 0,46. También se obtuvieron las palabras más relevantes para el modelo de regresión logística al momento de obtener la calificación asociada a una reseña. Estas fueron todas adjetivos:



Importancia de las palabras - Regresión logística

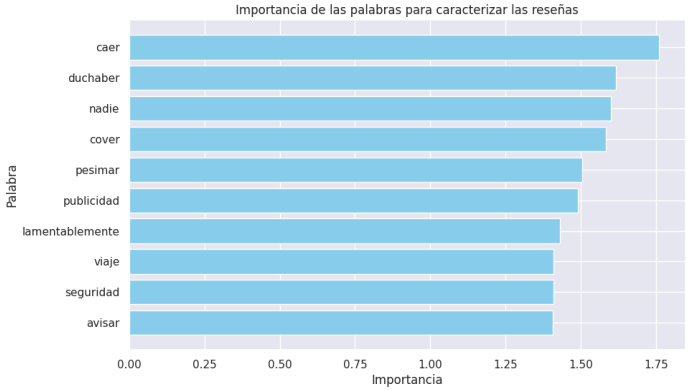
Support Vector Machine - Laura Restrepo

SVM es un algoritmo usado en tareas de clasificación que consiste en crear líneas divisorias entre grupos de datos. A partir de estas líneas, busca aquellas que más separación tengan entre ellas. De esta manera, puede ser usado para analizar textos, al encontrar características comunes y diferenciadoras. En una iteración inicial se crea un modelo sencillo para conocer las métricas y resultados iniciales. Se obtiene una exactitud, precisión, recall y score de f1 de aproximadamente 0.49 para todos. Esto da a entender que el rendimiento del modelo es moderado, y que existe espacio para mejora. Para esto, se observa el peso que tiene cada palabra en el resultado y, de este modo, conocer si se debe hacer un filtrado.

Importancia palabras iteración 1 SVM

Se obtiene que la mayoría de palabras son adjetivos, los cuales son importantes para distinguir el sentimiento que un turista tiene hacia el lugar visitado, pero no indican exactamente qué se debe mejorar. Por esto, en la segunda iteración se eliminarán los adjetivos para conocer cómo mejoran las métricas.

Una vez construido el modelo sin los adjetivos, se obtienen métricas de alrededor de 0.43. A pesar de que tengan un valor menor, los resultados obtenidos en la importancia de las palabras se alinean más a los objetivos de negocio, pues permiten distinguir de manera más clara aquellas características diferenciadoras.



Importancia palabras iteración 2 SVM

Resultados

Dentro de los resultados obtenidos en los modelos anteriores, se puede observar que las palabras más relevantes al momento de determinar el sentimiento de una reseña o, en este caso, su calificación asociada, son los adjetivos que contiene, ya que los adjetivos son cruciales para expresar opiniones y emociones en el lenguaje humano, lo que los convierte en indicadores clave para determinar el sentimiento de cada texto. Sin embargo, si se desea conocer información aparte de los adjetivos, las segundas iteraciones muestran resultados con palabras más alineadas al

objetivo del negocio. Ejemplo: palabras como ‘seguridad’, dan a entender que es una prioridad para los turistas. ‘Publicidad’ puede referirse a publicidad engañosa. Este tipo de palabras son bastante útiles pues les pueden decir a las cadenas hoteleras cómo pueden mejorar los sitios turísticos.

Mapa de actores relacionados con el producto de datos creado

A continuación se encuentra una descripción de los actores dentro de las cadenas hoteleras que pueden beneficiarse del resultado del modelo analítico planteado:

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Dueño	Usuario-cliente	Más ingresos y mejora de la reputación de la cadena hotelera.	Si el modelo no identifica correctamente las mejoras, se puede empeorar la reputación y perder ingresos.
Desarrollador inmobiliario	Proveedor	Un nuevo establecimiento estratégicamente ubicado puede atraer más huéspedes y así ingresos.	Construir un establecimiento lleva muchos costos, por lo que puede ser una inversión perdida.
Empleados del hotel	Proveedor	Al mejorar los sitios turísticos y atraer más clientes, se pueden generar más empleos y/o aumentar salarios.	Si el hotel empieza a invertir incorrectamente en los sitios turísticos pueden perder su empleo.
Huésped (turista)	Beneficiado	Tener un hotel cercano a buenos sitios turísticos.	Si el modelo no funciona, puede
Inversionistas	Financiador	Si el modelo funciona correctamente y se invierte dinero en las oportunidades identificadas, puede recuperar su inversión.	Si el modelo no funciona y destina su dinero incorrectamente, puede perder su inversión.
Sitio turísticos	Facilitador	Un buen modelo ayudaría a identificar mejoras y atraer más turistas.	Si no se identifican las mejoras correctamente, puede perder turistas y, así, ingresos.

Trabajo en equipo

Para este trabajo, los roles definidos entre los miembros fueron:

- Gabriel Dicelis - Líder de analítica
- Juan Pablo Martinez - Líder de datos
- Laura Restrepo - Líder de proyecto y líder de negocio

También se definió el siguiente cronograma:

Fecha	Responsabilidad	Tarea	Tiempo dedicado
Marzo 27	Grupal	Revisar el enunciado	0.5 horas
Marzo 29	Grupal	Explorar los datos	2 horas
Marzo 30	Grupal	Empezar documento	2 horas
Marzo 31	Grupal	Verificar, limpiar y preparar los datos	3 horas
Abril 1	Individual	Crear algoritmo	4 horas
Abril 2	Grupal e individual	Avanzar documento	2 horas
Abril 3	Grupal	Seleccionar algoritmo e implementarlo	2 horas
Abril 4	Grupal	Terminar documento	2 horas
Abril 7	Grupal	Grabar video	0.5 horas

Por último, estas fueron los retos identificados y las estrategias usadas para mejorarlos:

- Encontrar algoritmos para trabajar. Para ello se optó por hacer una búsqueda grupal y acordar en grupo aquellos que se consideran mejores.
- Identificar estrategias para mejorar las métricas. Para ello se hizo una consulta en grupo para proporcionar apoyo en dichas estrategias.

Sustentación

Como el algoritmo con mejor score de f1 fue SVM, se selecciona este como el modelo a usar para predecir la calificación de nuevas reseñas. Para ello, se destinó una parte del notebook dedicada a crear, a partir del modelo, un archivo .csv con una nueva columna que incluyera la calificación dada.[3]

[3] Este archivo .csv puede ser encontrado en los anexo

Referencias

1. Mathworks. (s.f.) *Introducción a Support Vector Machine (SVM)*. Recuperado de https://la.mathworks.com/discovery/support-vector-machine.htm
2. Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing* (3.ª ed.). https://web.stanford.edu/~jurafsky/slp3/

Anexos

- Repositorio GitHub: https://github.com/Laurarestrepo03/Proyecto-1-BI
- Conjunto de datos limpio y preparado: https://drive.google.com/file/d/1AEvGCopVU5WpWvvAu-b3Kr6f6OVzMd-a/view?usp=drive_link
- Conjunto de datos de sustentación: https://drive.google.com/file/d/1B7DQ_pwjLk9nU2Vpb5Ox6dJjfgYhm4Es/view?usp=drive_link
- Video de resultados: https://youtu.be/Q6bw4kUWiYg