PROYECTO 1- ETAPA 1

CONSTRUCCIÓN DE MODELOS DE ANALÍTICA DE TEXTOS

TURISMO DE LOS ALPES

Daniel Acevedo - 201910941

Ingrith Barbosa - 201712085

Daniela Camacho – 202110974

ISIS3301

INTELIGENCIA DE NEGOCIOS

UNIVERSIDAD DE LOS ANDES

2024-10

Tabla de contenido

1. Entendimiento del negocio y enfoque analítico

Oportunidad/problema Negocio	Realizar una clasificación de las reseñas de los turistas para actuar con diversos establecimientos relacionados con el sector turístico. Esto busca analizar y categorizar las opiniones y valoraciones de los usuarios sobre diferentes negocios turísticos, como hoteles, restaurantes, agencias de viajes, entre otros. La finalidad de esta clasificación es identificar tendencias, áreas de mejora y posibles problemas que puedan afectar la experiencia de los turistas en Colombia, con el objetivo de implementar medidas correctivas y estrategias que contribuyan al desarrollo y fortalecimiento del turismo en el país.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Para responder al requerimiento de las organizaciones interesadas se utilizará una categorización de textos utilizando un modelo de aprendizaje supervisado. Este modelo deberá procesar el texto ingresado y a partir de este clasificar la reseña en una calificación de 1 a 5. Esta clasificación se logra después de realizar una preparación detallada de los datos y con el resultado de la preparación predecir una calificación. Los algoritmos a utilizar son: 1. Naive Bayes 2. Máquinas de vector de soporte 3. Regresión Logística
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El Ministerio de Comercio, Industria y Turismo de Colombia se beneficia de esta clasificación de reseñas porque le brinda una perspectiva directa de lo que piensan los turistas sobre los servicios y establecimientos turísticos del país. Con esta información, pueden identificar rápidamente dónde están los problemas y dónde se destacan, lo que les permite enfocar mejor sus esfuerzos en mejorar la experiencia turística en Colombia.
Contacto con experto externo al proyecto y detalles de la planeación	Alejandro Mejía – 201822170 Pablo Echeverri- 201920855 Reuniones: 1. Reunión de lanzamiento: 19 de marzo 2. Reunión de seguimiento: 4 de abril 3. Reunión de finalización: 6 de abril

2. Entendimiento y preparación de datos

Para el entendimiento de los datos, se identificó para cada reseña la palabra más pequeña y grande. Además de esto la palabra más común dentro de la reseña. Sin embargo, se identificó que la palabra más repetida en las reseñas son stopwords (es, de, el, y, que…), por lo que no dan ningún tipo de información. Se verificaron los idiomas de cada reseña y se identificó que no todas están en español. Sin embargo, se observó que algunas reseñas que no

están en español ni en inglés están mal clasificadas por la librería. A pesar de estar en español, la librería las identifica como si estuvieran en otro idioma. Por último, se verifico los valores de la calificación de la reseña y se evidencia que todos están en el rango especificado de 1-5.

A partir de esto, se realizó la preparación de datos que consta de tres pasos generales: Traducción, tokenización y normalización.

La traducción se realizó a aquellas reseñas que estaban en ingles por medio de la librería googletrans. Esto para que todas las reseñas estén en el mismo idioma.

La tokenización consta en dividir los textos en palabras individuales por lo que cada token va a representar una palabra de la reseña. Al dividir el texto en tokens, se simplifica la tarea de limpieza y normalización, ya que se puede aplicar cada paso de preprocesamiento de manera individual a cada token. Además, la tokenización facilita la extracción de características y la representación del texto de una manera que los algoritmos de aprendizaje automático puedan entender y procesar de manera efectiva.

En el proceso de normalización lingüística, se aplican una serie de transformaciones a cada token del texto:

1. Convertir el texto a minúscula.
2. Eliminar caracteres no alfabéticos.
3. Eliminar espacios en blanco adicionales.
4. Eliminar stopwords.
5. Lematización: Convertir las palabras a su forma base o raíz.
6. Normalización de números: Representar los números utilizando palabras en lugar de cifras numéricas.

Finalmente, se realizó bag of word para representar las calificaciones de forma numérica y que los algoritmos pudieran trabajar con los datos correctamente.

3. Modelado y evaluación

Se escogieron los siguientes algoritmos:

3.1 Support Vector Machine (SVM)

El algoritmo SVM (Support Vector Machine) es un método de aprendizaje supervisado que se utiliza principalmente para clasificación, aunque también puede aplicarse a problemas de regresión. Su objetivo es encontrar un hiperplano en un espacio de características que maximice el margen entre las clases de datos, donde el margen es la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase, conocidos como vectores de soporte. SVM funciona encontrando el hiperplano óptimo que separa las clases de datos utilizando una función de pérdida, generalmente la pérdida de bisagra, que penaliza las clasificaciones incorrectas. Además, SVM puede manejar datos no linealmente separables mediante el uso del truco del kernel,

que mapea los datos a un espacio de características de mayor dimensión donde son linealmente separables. Esto hace que SVM sea un algoritmo eficaz para la clasificación en una amplia variedad de problemas, ya que puede manejar tanto datos linealmente como no linealmente separables.

Seleccionamos este algoritmo debido a su capacidad para manejar eficientemente datos no linealmente separables, su eficiencia en espacios de características de alta dimensión, su habilidad para generalizar y evitar el sobreajuste, y su capacidad para identificar características relevantes a través de la interpretación de vectores de soporte. Adicionalmente, este algoritmo resulta útil para determinar qué palabras se utilizan para realizar la clasificación, pues se lleva a cabo un análisis de características importantes, examinando los coeficientes asociados a cada característica o las palabras más frecuentes entre los vectores de soporte, lo que permite identificar qué términos son más influyentes en la clasificación de las reseñas de turistas.

3.2 Regresión logística

La regresión logística es un algoritmo que resuelve problemas de clasificación. Estima la probabilidad de que una instancia pertenezca a una clase utilizando una función logística. Esta función, también conocida como función sigmoide, toma una combinación lineal de las características de entrada y los pesos asociados, y la transforma en un valor entre 0 y 1, que representa la probabilidad de pertenencia a una clase positiva. Para predecir la clase de una instancia, se compara la probabilidad estimada con un umbral (generalmente 0.5). Si la probabilidad es mayor que el umbral, se clasifica como clase positiva; de lo contrario, se clasifica como clase negativa.

Se eligió la Regresión Logística para abordar este problema debido a su eficacia en la clasificación multiclase, como es el caso de la tarea en cuestión. Esta elección se justifica por la capacidad del algoritmo para manejar múltiples clases de forma eficiente y producir resultados que son fácilmente interpretables. Además, la Regresión Logística cuenta con técnicas de regularización integradas que ayudan a evitar el sobreajuste y mejorar la capacidad de generalización del modelo, especialmente en conjuntos de datos de texto con alta dimensionalidad. En resumen, la Regresión Logística proporciona una combinación de rendimiento, interpretabilidad y control de sobreajuste que la hace adecuada para esta tarea específica de clasificación de texto multiclase.

3.3 Naive Bayes

Naive Bayes es un algoritmo de aprendizaje supervisado basado en el teorema de Bayes. Se utiliza para la clasificación de datos. Su "naive" (ingenuo) viene de la suposición de independencia condicional entre las características, lo que significa que asume que todas las características son independientes entre sí dado el valor de la clase.

El algoritmo calcula la probabilidad de que una instancia pertenezca a una determinada clase basándose en la probabilidad condicional de las

características dadas las clases. Luego, clasifica la instancia en la clase con la probabilidad más alta.

Es simple, eficiente y funciona bien en conjuntos de datos con muchas características, como el análisis de texto. Es especialmente útil cuando se tienen datos de alta dimensionalidad como el bag of words generado a partir de reseñas turísticas.

Ahora, en cuanto a por qué se eligió Naive Bayes para este proyecto:

Se eligió Naive Bayes debido a su eficiencia y buen desempeño en la clasificación de texto, que es el tipo de datos que estás analizando con las reseñas turísticas. Su simplicidad y velocidad lo hacen adecuado para este tipo de tarea. Además, su capacidad para manejar características de alta dimensionalidad como las palabras en las reseñas lo hace ideal para el análisis de sentimientos en este contexto.

3.4 Modelo seleccionado

Seleccionamos el modelo de regresión logística ya que presenta los valores más altos de la métrica F1, que tiene en cuenta tanto la precisión como el recall, los cuales eran requerimientos importantes para las organizaciones interesadas. Las raíces de las palabras que el modelo utiliza para realizar la clasificación son: pesim, horribl, peor, impresion, histori, desastr, engañ, lament, excelent, rot. Como se puede observar, se presentan las raíces de palabras como horrible y excelente, palabras determinantes a la hora de calificar un lugar y darle un valor numérico. Se concluye finalmente que el algoritmo tiene en cuenta la incidencia de las estas palabras a la hora de clasificar el texto.

4. Resultados

4.1 SVM

Para observar los resultados de cada modelo realizado se generó una matriz de confusión, en ella se observa la cantidad de datos clasificados en cada clase. En la matriz del modelo SVM se puede observar que la mayoría de los datos clasificados se encuentran ubicados en la diagonal de esta, lo que es buen indicador ya que son los datos que fueron correctamente clasificados. Sin embargo, para algunas clases intermedias como 3 la cantidad de datos que se encuentran adyacentes a la diagonal son bastante grandes. Esto puede deberse a que, son datos difíciles de clasificar, pues al ser calificaciones intermedias tienen comentarios buenos y malos lo que puede confundir al modelo.

Para realizar la evaluación cuantitativa del modelo se observan las métricas de precision, recall y f1-score de la siguiente manera:

Test Report for SVM

	precision	recall	f1-score	support
1	0.42	0.49	0.45	161
2	0.33	0.33	0.33	227
3	0.35	0.38	0.37	299
4	0.42	0.38	0.40	401
5	0.61	0.59	0.60	487
accuracy			0.45	1575
macro avg	0.43	0.43	0.43	1575
weighted avg	0.45	0.45	0.45	1575

El modelo presenta una precisión (precision) que varía entre 33% y 61% y una sensibilidad (recall) que oscila entre 33% y 59%, teniendo menores valores las clases intermedias. Esto indica que para algunas clases como la 5 se clasifican correctamente una mayor cantidad de datos. Sin embargo, para la clase 2 y 3 se predicen correctamente una menor cantidad de datos, esto se explica cómo se mencionó anteriormente por ser clases intermedias. Estas métricas indican que el modelo clasifica mejor las clases borde, las reseñas muy buenas o malas. Sin embargo, le cuesta más clasificar las reseñas intermedias. Finalmente, la precisión general del modelo es del 45%, lo que sugiere que aproximadamente la mitad de las instancias se clasifican correctamente.

4.2 Regresión logística

Para el análisis de los resultados, como se mencionó anteriormente, se realizó una matriz de confusión con el fin de interpretar los resultados del modelo. Al observar la matriz, se evidencia que la mayoría de los datos se encuentran en la diagonal principal. Este patrón es alentador, ya que indica que la mayoría de los datos fueron etiquetados correctamente por el modelo.

Además, al analizar el eje X de la matriz, se observa que la etiqueta 3 es la que contiene la mayoría de los datos mal etiquetados. Esto sugiere que el modelo tuvo dificultades específicas al clasificar correctamente las reseñas asociadas con esa etiqueta.

Por otro lado, es importante destacar que las reseñas con una calificación de 5 fueron las que el modelo predijo con mayor precisión, lo que sugiere que el modelo tuvo un buen desempeño al clasificar las reseñas más positivas.

Ahora, las métricas que obtuvo el modelo fueron las siguientes:

Test Report for Logistic Regression

	precision	recall	f1-score	support
1	0.45	0.43	0.44	161
2	0.42	0.37	0.39	227
3	0.40	0.40	0.40	299
4	0.42	0.39	0.40	401
5	0.58	0.65	0.61	487
accuracy			0.47	1575
macro avg	0.45	0.45	0.45	1575
weighted avg	0.47	0.47	0.47	1575

Este informe revela que la precisión del modelo oscila entre el 40% y el 58%, el recall entre el 37% y el 65%, y el puntaje F1 entre el 39% y el 61%. Estos resultados reflejan que el desempeño del modelo varía significativamente según la clase. Específicamente, se observa que la clase con la menor precisión, recall y puntaje F1 es la clase 3, mientras que la clase 5 presenta los mejores resultados en términos de estas métricas.

Estos hallazgos son consistentes con lo observado en la matriz de confusión, donde se evidenció que la clase 3 fue la más problemática en términos de clasificación incorrecta. Por otro lado, las reseñas con una calificación de 5 fueron las que el modelo clasificó con mayor precisión, lo que sugiere una habilidad destacada del modelo para reconocer y clasificar las reseñas más positivas.

Con esto se puede decir que el modelo esta más entrenado para reconocer patrones asociados con reseñas extremadamente positivas o negativas, mientras que puede tener dificultades para identificar reseñas con calificaciones intermedias.

4.3 Naive Bayes

Para el análisis de resultados, al igual que los demás algoritmos se hizo una matriz de confusión que permitiera determinar el rendimiento del modelo. En este caso, la matriz de confusión indica que el modelo tiene un buen rendimiento, ya que la mayoría de las predicciones se encuentran en la diagonal principal. Esto significa que el modelo ha clasificado correctamente la mayoría de las instancias.

Sin embargo, también hay algunos errores de clasificación. Por ejemplo, el modelo ha confundido algunas instancias de la clase 1 con la clase 2, y viceversa. Además, el modelo ha clasificado algunas instancias de la clase 4 como pertenecientes a la clase 3.

Por otro lado, al determinar el accuracy, precisión, recall, f1-score y support se obtuvo los siguientes resultados:

Accuracy: 0.48634920634920636

	precision	recall	f1-score	support
1	0.64	0.23	0.34	161
2	0.42	0.46	0.44	227
3	0.38	0.35	0.36	299
4	0.42	0.49	0.45	401
5	0.62	0.67	0.64	487
accuracy			0.49	1575
macro avg	0.49	0.44	0.45	1575
weighted avg	0.50	0.49	0.48	1575

Los resultados de la evaluación del modelo Naive Bayes muestran que, si bien logra una precisión global del 49%, su desempeño varía entre las diferentes clases. Se observa una alta precisión para la clase 5 (62%) y menor para la clase 1 (64%), aunque con un recall bajo (23%). Esto sugiere que el modelo es mejor identificando reseñas positivas de alta puntuación, pero tiene dificultades para clasificar correctamente las de baja puntuación.

La sensibilidad promedio ponderada es del 49%, lo que indica que el modelo podría beneficiarse de ajustes adicionales para mejorar su capacidad de identificar las clases minoritarias y, así, lograr un balance óptimo entre precisión y recall en la clasificación de reseñas turísticas.

5. Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Departamento de turismo	Cliente/Usuario	Obtener información detallada sobre las opiniones y valoraciones de los turistas lo que permite identificar áreas de mejora y tendencias aprovechables en el sector turístico colombiano. Esto facilita la toma de decisiones informadas para implementar las medidas y estrategias necesarias	Dependencia excesiva del modelo de clasificación de las reseñas turísticas como única fuente de retroalimentación sobre la calidad de los servicios brindados. Lo anterior puede llevar a decisiones erróneas si no se consideran otros factores.
Agencia de viajes	Cliente/Usuario	Acceso a información detallada sobre la percepción de los turistas sobre los servicios ofrecidos, permitiendo así ajustar y personalizar ofertas para satisfacer más las necesitades y preferencias de los clientes	Dependencia excesiva de la clasificación automatizada, lo que podría llevar a una falta de comprensión de las necesidades individuales de los clientes.
Propietarios de hoteles	Cliente/Usuario	Identificación rápida de áreas de mejora y problemas potenciales que puedan afectar la experiencia de los huéspedes, permitiendo así,	Dependencia exclusiva del modelo de clasificación de reseñas turísticas para evaluar la satisfacción de los huéspedes, lo que podría descuidar otros aspectos

		tomar medidas correctivas de manera oportuna y eficiente	importantes como la atención al cliente y la calidad de las instalaciones
Restaurantes	Cliente/Usuario	Obtener información detallada sobre las preferencias y opiniones de los comensales, lo que les permite ajustar su oferta y mejorar la calidad de sus servicios para satisfacer mejor las expectativas de los clientes.	Dependencia excesiva de la clasificación automatizada de reseñas turísticas, lo que podría llevar a una falta de comprensión de las necesidades individuales de los clientes y a una pérdida de la capacidad de innovación y creatividad en la oferta gastronómica.
Proveedores de Servicios Turísticos	Proveedor	Ofrecimiento de herramientas analíticas avanzadas para comprender mejor las opiniones y valoraciones de los turistas, lo que les permite mejorar la calidad de sus servicios y diferenciarse en el mercado turístico.	Dependencia del modelo de clasificación de reseñas turísticas desarrollado por terceros, lo que podría generar preocupaciones sobre la confiabilidad y la adaptabilidad del modelo a las necesidades específicas de cada proveedor de servicios turísticos.

6. Trabajo en equipo

6.1 Roles

Líder de proyecto	Ingrith Barbosa
Líder de negocio	Daniel Acevedo
Líder de datos	Ingrith Barbosa
Líder de analítica	Daniela Camacho

6.2 Tareas realizadas

Ingrith		Daniela		Daniel	
Construcción y		Construcción y		Construcción y	

entendimiento del modelo	30 minutos	entendimiento del modelo	30 minutos	entendimiento del modelo	30 minutos
Análisis de resultados	30 minutos	Análisis de resultados	30 minutos	Análisis de resultados	30 minutos
Construcción pipeline y predicción de datos	30 minutos	Preparación de los datos	60 minutos	Revisión de cumplimiento de entregables y envío	20 minutos
Contribución documento	45 minutos	Contribución documento	60 minutos	Contribución documento	30 minutos
Contribución video	45 minutos	Contribución video	45 minutos	Contribución video	15 minutos